The Design of Summative Evaluations for the Employment Benefits and Support Measures (EBSM)

(For the Human Resources Development Canada (HRDC) Expert Panel on Summative Evaluation Design for the EBSM Program)

Prepared by:
Walter Nicholson

September 2001

SP-AH168-09-01E
(également disponible en français)

106-02331/

Table of Contents

Preface i

A. Defining the "Participant" in Employment Benefits and Support Measures (EBSM) Evaluations 1

B. Analysis Methods to be Used in the Evaluations 5

C. Sample Design 17

D. Outcome Specification and Measurement 25

E. Coordination with the medium term indicators (MTI) Project 29

F. Summary of Issues Requiring Further Consideration at Implementation 31

References 35

Annex A: Summative Evaluations of Employment Benefits and Support Measures and Labour Market Development Agreements — Possible Program Effects to be examined/measured 37

Preface

Human Resources Development Canada (HRDC) and its provincial and territorial partners recognize the importance of evaluation evidence in helping them assess the impacts of the Employment Benefits and Support Measures and the Labour Market Development Agreements (EBSM/LMDA). Provision for evaluation of these is found in both the Employment Insurance legislation and in the LMDAs.

Generally, the evaluation of employment interventions is a complex undertaking. In the case of the summative evaluations of the EBSM/LMDA, the complexity is compounded by a large diversity in communities and delivery models across the country, the number of individual evaluation studies to be undertaken, and the timing of each.

Evaluation studies involve a blend of creative thinking and in-depth knowledge of measurement techniques. This is so partly because they must combine knowledge of complicated labour market and economic theories, intricate models of program delivery and learning, elaborate sets of expected and unexpected potential impacts, and highly structured statistical models to estimate outcomes and impacts from partial lists of concepts that often cannot be measured directly.

By their very nature, EBSM/LMDA evaluations involve a large number of partners with a wide range of interests. They include policy makers, program managers, delivery officers, consultants, and citizens (such as program participants and taxpayers). These groups possess different levels of understanding of evaluation principles, standards, techniques and interpretations of the evidence. For the evaluations to succeed, it is essential to establish a high level of trust among all partners/interested parties. Therefore, it is important that the methodology be highly defensible and that it be presented in a lucid and fully transparent manner.

HRDC engaged an expert panel to develop a generic, objective, state-of-the-art framework for the evaluation of the EBSM/LMDAs. The framework was developed by the experts to provide a common and sound methodological foundation for the various evaluations. A guiding principle was that it be flexible enough to meet requirements for reports at the national, provincial, territorial and local levels. A key feature is that it is comprehensive and readily adaptable to specific operational circumstances and subject matter content. The proposed methodology is based on the latest theoretic and applied evaluation approaches and views and experiences of the experts with respect to their application.

The Panel's proposed framework is summarized in the body of this report. It should facilitate agreement with respect to methodology so that at the time of reporting the focus will be on interpretation of the evidence in terms of likely impacts of the EBSM/LMDAs. The body of this report is devoted almost exclusively to methods of measuring outcomes, and not to what ought to be measured. The proposed measurement framework can be applied to a wide range of outcomes (for example, those discussed in Annex A).

HRDC would like to thank Mr. Walter Nicholson, Mr. Jeff Smith, Mr. Chris O'Leary, Mr. Paul Decker, and Mr. Joe Michalski for their excellent work.

HRDC offers this report in the spirit of openness to advance the understanding and facilitate discussions of evaluation methodology, and to ensure the best set of possible summative evaluations of the EBSM.

This report presents a general framework that can form the core of any well-designed evaluation of the Employment Benefits and Support Measures (EBSM)[1]. It identifies questions that should be addressed, offers assessments of strengths and weaknesses of various approaches for measurement, and highlights issues toward which more research effort might be directed[2].

Even though the paper does go into considerable depth on some matters, it should not be read as a definitive detailed plan for any given evaluation as such. As discussed in the paper, the framework will have to be adapted (modified and extended) to reflect features that may be unique to the particular requirements of any given evaluation. For example, any final design for an evaluation must ultimately depend on the goals set for that evaluation, the nature of the programmatic interventions to be studied, the quality of the data available, and on such practical considerations as time and resource availability.

An important consideration for the evaluation of the EBSM concerns the reporting and ultimate perceived validity of evaluation evidence. Although this paper attempts to describe the "state-of-the-art" with respect to evaluation methodology, it does not pretend to be the final word on what a "good" evaluation is, nor how the evidence issuing from such methodology should be used. Assessing the overall validity of the evidence relies, in large measure, on: (1) An *a priori* understanding of the applicability and limitations of the general approach being used (e.g., limitations of quasi-experimental designs); and (2) An *ex post* appraisal of the quality of the data collected and of robustness of the estimation strategies used. This is especially important in the context of the EBSM evaluations where it is likely that a major component of most evaluations will be to provide evidence to assist in the assessment of the "incremental effects" of the interventions being offered. While well-understood and organized data can contribute a great deal, the problems associated with deriving reliable estimates of such incremental effects in non-experimental contexts are pervasive. In many cases it may simply prove to be impossible to provide valid results based solely on the available data. Hence, planned evaluations must include procedures for the assessment of strengths and weaknesses together with specification of a clear process for the integration of results and their dissemination to wider audiences.

These cautionary notes are an integral part of the rationale for the framework articulated in this paper and are re-enforced throughout. A proper understanding of the Panel's suggestions requires that they be viewed within this cautionary context. Recognizing the potential strengths and weaknesses associated with various approaches in the development of the summative evaluations will be instrumental in the proper interpretation and presentation of the evidence gathered and reported in future.

---

1 EBSM stands for "Employment Benefits and Support Measures". This terminology is used at the national level, but regional terms for the program may vary.

2 This report assembles, organizes and extends the contents of a series of papers that summarize the views of the EBSM Expert Panel. The views of the panel were obtained during a two-day meeting (March 8 and 9, 2001 – see Nicholson 2001) and in sundry and ad hoc ongoing discussions.

The paper is divided into five analytical sections that focus on the following general conceptual points:

A. Definition of the EBSM Program Participation
B. Analysis Methods (including Comparison Group Selection)
C. Sample Design
D. Outcome Measurement
E. Integration with the medium term indicators (MTI) Project.

A final section (F) summarizes some of the issues raised here that should be kept in mind when designing quasi-experimental evaluations and includes some guidelines regarding implementation of the EBSM evaluations.

A. Defining the "Participant" in Employment Benefits and Supports Measures (EBSM) Evaluations

In order to structure a clear analysis of the likely effect of the interventions that constitute the EBSM program, one needs to develop a precise definition of what that program is and how "participants" in it are to be identified. Two general factors guided the Panel in its considerations of these issues: (1) The program and its constituent interventions should be defined in ways that best reflect how services are actually delivered; and (2) Participation should be defined in a way that both reflects individuals' experiences and does not prejudge outcomes. Given these considerations, the panel made the following observations and suggestions:

1. A specification under which program participants are defined based on Action Plan start dates seems to best meet evaluation goals. The panel believed that the Action Plan concept best reflects the overall "case management" approach embodied in the EBSM program. Although participants actually experience quite different specific interventions, most share a common process of entry into the program together with a conscious direction to services that are deemed to be most beneficial. Hence, the Action Plan notion (although it may not be explicitly used in all regions) seemed to correspond most directly with what policy-makers believe the "program" to be.

The use of start dates as a way of identifying program participants appeared to the Panel to offer a number of advantages. First, it seems likely that start dates are defined in a relatively unambiguous way for all program participants. The alternative of focusing on participants who are ending the program during a specific period seemed more problematic, in part because Action Plan end dates are often arbitrarily defined. A second reason for preferring a start date definition of participation is that it seems likely that such a definition would better match-up with other published data, such as those in the Monitoring and Assessment Reports. Finally, opting for a start date definition of participation seemed to provide a conceptually clearer break between pre-program activities and "outcomes" because, for some purposes, everything that occurs after the start date can be regarded as "outcomes". That is, the in-program period is naturally included so that opportunity costs associated with program participation can be included as part of an overall assessment.

Of course, the Panel recognized that some regions do not use a case-management/Action Plan approach to delivering EBSM interventions. They also recognized that basing a definition of participation on start dates poses conceptual problems in cases of multiple interventions or when there is policy interest in differentiating between in-program and post-program outcomes. Although the Panel does not believe that these potential disadvantages negate the value of their suggested approach (which seems flexible enough

to accommodate a wide spectrum of actual programmatic procedures[3]), it did believe that designers of specific evaluations should explore these issues again in the light of actual program processes and data availability.

2. In general, evaluations should focus on individuals who have participated in an "Employment Benefit". The four employment benefits ("EB's" — that is, Targeted Wage Subsidies, Self Employment, Job Creation Partnerships, and Skills Development) constitute the core of EBSM offerings. They are also the most costly of the interventions offered on a per participant basis. Therefore, the Panel believed that the evaluations should focus on participants in these interventions. It seems likely that regions will wish to assess the impacts of each of these interventions separately. That desire poses special problems for the design of the evaluations. These are discussed in detail in Section C below. There we show that optimal designs for achieving the goal of assessing individual interventions may differ depending on the specific interests of regional policy-makers. The Panel also noted that some consideration might be given to including a separate group of participants in Support Measures only in some evaluations. That possibility is discussed below.

3. "Participation" requires a precise definition — perhaps defined by funding (if feasible). The Panel expressed concern that a number of EBSM clients may have start dates for specific Employment Benefits (or for Action Plans generally) but spend no actual time in any of the programs. Assuming that such "no shows" are relatively common (though some data should be collected on the issue), the panel believed that there should be some minimal measure of actual participation in an intervention. One possibility would be to base the participation definition on observed funding for an individual in the named intervention. Whether individual-specific funding data for each of the EB interventions are available is a question that requires further research. It seems likely that funding-based definitions of participation would differ among the regions because ways in which such data are collected would also vary.

4. Program completion is not a good criterion for membership in the participant sample. Although it might be argued that it is "only fair" to evaluate EBSM based on program completers, the Panel believed that such an approach would be inappropriate. Completion of an EB is an outcome of interest in its own right. It should not be a requirement for membership in the participant sample.

5. A separate cell for Employment Assistance Services (EAS)-only clients should be considered in some regions. The EBSM program allocates roughly one-third of its budget to Support Measures and these should be evaluated in their own right. Prior research has shown that relatively modest employment interventions can have the most

---

3 For example, in regions without an explicit case management/Action Plan approach, it still would be possible to simulate Action Plan start and end dates using start and end dates of interventions. Further study of the precise way of accomplishing this simulation will require in-depth knowledge of local policies and delivery practices. It is possible that definitions will have to vary across regions if the simulated plans are to reflect accurately the ways in which programs actually operate. This is an issue that the joint evaluation committees and Human Resources Development Canada (HRDC) need to consider carefully when finalizing their provincial/territorial evaluation designs and reporting at the national level on results.

cost-effective impacts (see Section C). Hence, the Panel believed that dropping participants with only an SM intervention from the analysis runs the danger of missing quite a bit of the value of the overall EBSM program. That is especially true in regions that place a great deal of emphasis on Support Measures. It also seems possible that including an EAS-only sample cell might aid in estimating the impact of the EB interventions themselves. This is so because an EAS-only sample may, in some circumstances, prove to be a good comparison group for specific EB interventions. Many other evaluations have adopted a tiered approach to defining treatments in which more complex treatments represent "add-ons" to simpler ones. In some locations the EBSM may in fact operate in this tiered way. A few further thoughts on potential roles for an EAS-only treatment cell are discussed in Sections B and C; the Panel believes that most evaluations should consider this possibility.

6. Apprentices should be the topic of a separate study. Although the provision of services to apprentices is an important component of the EBSM program, the Panel believed that the methodological issues that must be addressed to study this group adequately would require a separate research agenda. Some of the major issues that would necessarily arise in such a study would likely include: (1) How should apprentices' spells of program participation be defined? (2) How is a comparison group to be selected for apprentices — is it possible to identify individuals with a similar degree of "job attachment"? And (3) How should outcomes for apprentices be defined? Because the potential answers to all of these questions do not fit neatly into topics that have been studied in the more general employment and training literature, the Panel believed that simply adding an apprenticeship "treatment" into the overall summative evaluation design would yield little in the way of valuable information. Such a move would also detract from other evaluation goals by absorbing valuable study resources.

B. Analysis Methods to be Used in the Evaluations

The Panel recognized that assessing the incremental effects of Employment Benefits and Support Measures (EBSM) interventions is an appropriate goal.[4] Doing so in a non-experimental setting (i.e., using quasi-experimental designs), however, poses major challenges. The methodological issues associated with deriving such "incremental effects" in the form of definitive estimates of the effect of an intervention on participant outcomes using comparison groups are profound. In many cases, based on available data, it may simply be impossible to provide definitive evidence regarding "true" impacts. In such circumstances, results should be cast in the form of evidence to be used as "indicators of potential impacts", under certain clearly specified conditions. Any extension of the results beyond those conditions should be subjected to careful scrutiny and critical analysis.

The issues may be illustrated as follows: a goal of a summative evaluation is to obtain "consistent" and relatively "precise" estimates of the outcomes for participants and comparison groups. That is, the methodology should yield estimates that, if samples were very large, would converge to the true (population) values. And the actual estimates should have a small sampling variability so that the estimates made have only small probabilities of being far from the true values. Indeed, managers often expect that the evaluation can provide consistent and precise estimates of the "incremental effect" (i.e., the effect of the program on outcomes relative to what would have happened in the absence of the intervention). The Panel recognized the difficulty of achieving these goals using a quasi-experimental design. Inherent in such a design is the basically unanswerable question of how similar comparison groups really are, especially when differences between groups may be embedded in characteristics not directly observable in the evaluation. The Panel believed that it is not possible to specify on *a priori* grounds one "best" approach that will optimally provide proper estimates in all circumstances. Hence, the Panel recommended that evaluators take a broad-based and varied approach, exploring a variety of methodologies both in the planning phase and in the analysis phase of their work. It also strongly believed that the approaches taken should be carefully documented and critically compared and that evaluators should be encouraged to suggest a best approach.

1. Potential Measurement Strategies

The Panel believed that a number of approaches to measuring the outcomes of participants and non-participants in EBSM interventions seem feasible for the summative evaluations.

---

4 Other goals that may be important to some policy-makers include: (1) Obtaining good data on the characteristics of program participants; (2) Studying the process through which EBSM interventions are delivered; and (3) measuring the "gross" post-program experiences of EBSM participants. Although most of this report (and the Panel's deliberations) focused on obtaining incremental estimates, these other goals may be quite important in specific regions and the goals will therefore play important roles in those regions' evaluations designs.

To ensure that all possibilities were considered, it developed a rather exhaustive list of the possibilities. What follows is a listing of those possibilities with some critical comments on each.

a. **Random Assignment ("Experimental") Methods:** Random assignment remains the "gold standard" in labour market evaluations. The procedure guarantees consistent and efficient estimation of the average effect of a treatment on the treated[5] and has become the standard of comparison for all other approaches (see, for example, Lalonde 1986 and Smith and Todd 2001). Because of these advantages, the Panel strongly believed that the possibility for using random assignment in some form in the summative evaluations should be considered. Of course, the Panel recognized that the primary objection to random assignment is that it conflicts with the fundamental universal access goal of all EBSM programs. To the extent that individuals selected for a control group would be barred from EBSM program participation (at least for a time), this denial of service would create an irreconcilable difference between evaluation needs and program philosophy. The usual solution to such conflicts is to design treatments in such a way that they represent an enhancement over what is normally available in the belief that denial of enhancements is not so objectionable on philosophical grounds. If funding for specific types of interventions is limited, such interventions themselves might be considered "enhancements" so random assignment in such situations may also be supportable. Other constraints (say, on program capacity) may also provide some basis for structuring random assignment evaluations in ways that do not conflict with universal program eligibility.

   The Panel generally concluded, based on evidence from experiences with the EBSM program in the field, that such options are not common within the program as currently constituted, however. Still, the Panel felt that evaluators should always investigate the possibilities for structuring a random assignment evaluation first before proceeding to second-best solutions. Evaluators should also report on where random assignment evaluations might be most helpful in clarifying ambiguous evidence upon which assessments of impacts may be claimed. This might serve to highlight possible ways in which random assignment might be most effectively used to supplement the other research initiatives directed toward the EBSM program.

   Random assignment should also be considered for purposes other than the forming of traditional treatment-control designs. For example, a variety of increasingly extensive program enhancements could be randomly assigned on top of a universal service option. Or random assignment could be used to create "instruments" that help to identify program participation in otherwise non-experimental evaluations. This could be accomplished by, say, imposing additional counseling requirements on a random group of EI recipients and this might generate random variation in

[5] This statement would have to be modified if the experimental treatment also affected individuals in the control group (say by placing participants first in job queues). For a discussion of more complex questions about what can legitimately be inferred from random assignment experiments see Heckman, Lalonde, and Smith 1999.

training receipt. A similar goal could be achieved through the use of randomly assigned cash bonuses. The main point is that random assignment could be used in a variety of innovative ways to improve the quality of the EBSM evaluations. Because of the additional validity that can often be obtained through random assignment, the panel would enthusiastically support such innovative efforts.

b. **Non-Experimental Methods:** A wide variety of non-experimental methodologies have been used for evaluating labour market programs when random assignment is infeasible. Most of these utilize some form of comparison group in the hope that experiences of comparison[6] group members faithfully replicate what would have happened to program participants had they not been in the program. We begin with a relatively simple listing of the possibilities. This is followed, in the next subsection, with a more detailed discussion of how comparison group methods interact with measurement methods in determining whether consistent estimates of outcomes can be achieved. Because, as we show, differences in performance of the methods depend in part on whether participants and comparison group members differ along dimensions that are observed or unobserved, we use this distinction to illustrate the approaches.

i. **Methods that Control for observable differences.** These methods are easy to implement, but, because they do not control for unobservable differences in the determinants of program participation, they remain suspect in their ability to provide consistent impact estimates[7]. Still, some literature suggests that matching strategies can be quite successful (Rosenbaum and Rubin 1983, Dehejia and Wahba 1998,1999) in generating useful comparison groups. So the Panel strongly suggested that they be considered as one method of measuring outcomes in the EBSM evaluations.

- **Comparison of means adjusted by Ordinary Least Squares (OLS).** These estimates generally are used to provide a "first cut" at the data. They do provide relatively efficient estimates that control (in a linear way) for observable differences between participants and comparison group members. The presentation of such results can help to clarify the nature of the data, but cannot be taken as definitive estimates of impacts. This is primarily because the approach provides no protection against the influence on outcomes of possible unmeasured differences such as those that relate to unobservable factors.

- **OLS with lagged outcome variables.** These estimates share similar problems to simple OLS estimates. In some cases controlling for lagged outcomes may improve matters by providing a partial control on unobserved differences between

6 Use of a before-after methodology does not require an explicit comparison group. Extrapolations from past behavior play that role instead.

7 Inconsistency can arise, for example, when an unmeasured variable (such as "motivation") affects both program participation and labour market outcomes. If more motivated individuals are more likely to participate and also have favourable labour market outcomes in the absence of participation, it will appear as if participation in the program "caused" such favourable outcomes.

participants and comparison group members[8]. But the use of lagged outcome variables can also introduce biases of unknown direction and resulting estimates can be very sensitive to precisely how the lagged variables are specified.

- **Matching Methods.** A variety of matching strategies might be employed in the EBSM summative evaluations[9]. As we discuss in the next section, these could be based only on administrative data or on some combination of administrative and survey data. While matching does not directly control for unmeasured differences between participants and comparison group members, it may approximately do so if the unobservable factors are correlated with the variables used for the matching[10]. Adoption of matching procedures would, as we show in Section C, have consequences for the sample allocation in the evaluations — primarily by increasing the size of comparison groups to allow for the possibility that some comparison group members included in the initial sample would not provide a good match to any participant. Adoption of matching strategies would also pose logistical problems in the collection of survey data for the evaluation and these also are discussed in Section C.

Two general approaches to matching have been employed in the literature:

- **Matching on Characteristics.** These procedures use either exact matching within cells based on participant characteristics or some sort of distance algorithm that matches comparison cases[11] to specific participant cases using observable characteristics. Estimates can differ widely depending on which specific characteristics are used in the matching routines. In some cases researchers have sought to simulate the uncertainties involved in exact matching by utilizing several different implementations of the matching algorithms.

- **Probabilistic Matching:** This process (pioneered by Rosenbaum and Rubin 1983) uses matching based on predicted "propensities" to participate in a program. It requires a first stage estimation of participation probabilities based on observable characteristics. The procedure may work well if two additional[12]

---

8 For example, since "motivation" affected past as well as future labour market outcomes, controlling for past outcomes does provide some (imperfect) measure of motivation.

9 In general matching methods might be considered superior to OLS procedures for controlling for observable variables because this method does not impose the linearity assumption implicit in OLS.

10 Technically, the data used for matching should be sufficient to ensure that participation in the program is unrelated to untreated outcomes conditional on the variables used in the matching.

11 For both matching approaches, comparison cases are chosen "with replacement" — that is, a comparison case may sometimes be the closest match for two or more participants. The efficiency loss from such double use is generally believed to be more than balanced by having better matches. It should be noted, however, that this method would be inappropriate for comparison group pools that are so small in relation to the treatment group that they would yield a large number of multiple draws (e.g., the pool for potential matches should not be smaller than two or three times the size of the participant group). More generally, matching procedures will not work well if the support set of characteristics differs substantially between comparison and participant groups.

12 That is, in addition to the general assumption required for all matching procedures that program participation be independent of untreated outcomes conditional on the observed variables.

assumptions are met: (1) if participation can be predicted "very accurately;" and (2) if the distribution of predicted probabilities are "similar enough" between participant and comparison groups. In actual practice probabilistic matching can also be somewhat difficult to implement because the determinants of participation are not well known and differing specifications of the participation equation may yield rather different matches. (Researchers should carefully list, explain, and provide the rationale for the criteria underlying the assessment of the terms "very accurately" and "similar enough.")

ii. Methods that Control for Unobservable Differences

These methods also proceed from assumptions about the nature of unobservable differences between groups of participants and non-participants. Then, on the basis of these assumptions, the methods attempt to account for and eliminate these *a priori* differences through subtraction, complex modeling, or careful selection of comparison groups on the basis of detailed characteristics. The validity of the estimates depends on the quality of the data availability, the extent to which the assumptions about unobservable factors hold, and the robustness of the techniques. It is important, therefore, that evaluators, in their assessments of the impacts of the EBSM, take into account deviations from the underlying assumptions and their potential effects on the corresponding estimates.

Specific approaches that attempt to control for unobservable factors include:

- **Difference-in-difference methods.** These methods are based on the crucial assumption that unobservable differences among individuals are constant for each individual over time — hence they drop out in the individual's data upon differencing on observables. With sufficient pre- or post-program observations, this assumption is, in principle, testable by using alternative time periods to compute before-and-after differences. If unobservable change over time or are affected by program participation, the difference-in-difference methodology will not yield consistent estimates, however.

- **Heckman/IV Methods.** These methods rely on the existence of an "instrumental variable" (IV) that must meet two criteria: (1) independence from the outcome being measured; and (2) significant predictive ability in explaining program participation decisions. Existence of "good" instruments is relatively rare, so evaluators adopting these techniques should be cautious. In some cases instruments can be generated within the confines of an evaluation[13]. The Panel believed that, when presenting results for these methods, evaluators should clearly specify what instrument was used and provide specification tests to evaluate whether the variable meets the necessary criteria. Evaluators should also consider the possibility of generating instrumental variables, when feasible.

---

13 For example, information on staff assessments of participant and non-participant suitability for a program may serve as such an instrumental variable.

2. Comparison Groups, Estimation Methods, and Consistency

The techniques described in the previous subsection have often been used in various ways in evaluations. In some cases, one procedure is adopted on an *a priori* basis and that method is the only one used. In other cases researchers have adopted an eclectic approach in which they may do some matching, a bit of regression analysis, and, for good measure, throw in some IV techniques. In general, the Panel feels that neither of these approaches is ideal. The selection of a single analytical approach to an evaluation may often prove inadequate once the data are examined in detail. But the adoption of a "kitchen sink" approach may lead to the adoption of techniques that are incompatible with each other. What seems required instead is some detailed rationale for any of the approaches chosen in a particular evaluation accompanied by an *ex post* appraisal of the likely validity of the approaches taken.

To aid in appraising issues related to the selection of analytical approaches to the evaluations, Table 1 explores the interaction between comparison group choice and estimation methods in some detail. The table considers three specific comparison group possibilities according to how these members are to be matched[14] to participant group members:

1. Matching on a limited set of administrative variables (for example Employment Insurance (EI) data only). These variables are termed V1. Virtually all designs will utilize this method of matching—if only to align dates at which recipients' benefit periods start;

2. Matching on V1 and additional administrative variables (V2 — which ideally would include earnings and income tax histories from Canada Customs and Revenue Agency (CCRA) because these are potentially the most informative administrative data). This would amount to fairly extensive matching on administrative data before any survey were undertaken; and

3. Matching on V1, V2, and additional variables that are only available from surveys, V3 (such as data on recent family earnings or on the process by which individuals entered/did not enter the EBSM program).

Each of these comparison group methods is related to four potential approaches that might be used to generate actual estimates of the differences in outcomes of interest between participant and non-participant groups:

1. Differences of Means;
2. OLS Adjusted Difference in Means;
3. Difference-in Differences (with OLS adjustments); and
4. Instrumental Variable (IV) Adjustment (including "Heckman procedures").

[14] This matching need not necessarily be pairwise, but could also include methods in which more than one non-participant is used to construct a "comparison" for each participant (see Heckman, Ichimura, and Todd, 1997).

TABLE 1
Consistency of Estimators in Various Circumstances

Comparison Group	Consistency[15] if Participant/Comparison Groups Differ in			
	V1 only	V1 or V2	V1orV2orV3	V1orV2orV3orU[16]
Simple Difference of Means				
V1 Match	Yes	No	No	No
V1,V2 Match	Yes	Yes	No	No
V1-V3 Match	Yes	Yes	Yes	No
Regression Adjusted Differences in Means* **				
V1 Match	Yes	Yes*	Yes*	No
V1,V2 Match	Yes	Yes	Yes*	No
V1-V3 Match	Yes	Yes	Yes	No
Difference-in-Difference Estimates				
V1 Match	Yes	Yes*	Yes*	Yes — If U Time Invariant. No otherwise
V1,V2 Match	Yes	Yes	Yes*	Yes — If U Time Invariant. No otherwise
V1-V3 Match	Yes	Yes	Yes	Yes — If U Time Invariant. No otherwise

* OLS adjustment will yield consistent estimates in these cases, though these will be less efficient than would be estimates based on matching using all measurable variables. Notice also that matching on all available variables would generally yield consistent estimates whereas OLS adjustment might not because of the linearity assumption implicit in OLS.

** Adoption of extensive matching algorithms may affect the efficiency of IV estimation methods.

*** It is assumed that OLS regressions would include all three types of measurable variables whereas matching would be only on the basis of the sets of variables identified in the table.

15 In the sense of statistical consistency as described earlier (if samples were very large, the estimates would converge to the true population values).

16 U means unobservable — these are variables that potentially affect both outcomes and program participation but not observed in either the administrative data or in the survey (e.g. motivation).

TABLE 1 (continued)
Consistency of Estimators in Various Circumstances

Comparison Group	Consistency if Participant/Comparison Groups Differ in			
	V1 only	V1 or V2	V1orV2orV3	V1orV2orV3orU
IV Estimators				
V1 Match	Yes**	Yes**	Yes**	Yes if instrument is good and IV procedure not compromised by matching[17]
V1,V2 Match	Yes**	Yes**	Yes**	Yes if instrument is good and IV procedure not compromised by matching
V1-V3 Match	Yes**	Yes**	Yes**	Yes if instrument is good and IV procedure not compromised by matching

* OLS adjustment will yield consistent estimates in these cases, though these will be less efficient than would be estimates based on matching using all measurable variables. Notice also that matching on all available variables would generally yield consistent estimates whereas OLS adjustment might not because of the linearity assumption implicit in OLS.
** Adoption of extensive matching algorithms may affect the efficiency of IV estimation methods.
*** It is assumed that OLS regressions would include all three types of measurable variables whereas matching would be only on the basis of the sets of variables identified in the table.

In order to understand this table, consider two examples. First, suppose that participants and comparison group members differ only along dimensions that are easily measured in the complete administrative data (that is, they differ only in V1 and V2 measures). In this case the third column of the table — labeled "V1 or V2"—is the relevant situation and the entries in Table 1 show that virtually all of the estimation procedures would work quite well no matter how the samples are matched[18]. In this case a first best solution would be to match on both V1 and V2 although regression-adjusted means with relatively modest (V1) matching might do equally well. Alternatively, in the more realistic case where participant and comparison group members differ along unmeasured dimensions (labeled "U"), the estimation procedures are quite varied in their performance. One approach that has been suggested at times, for example, is to use matching from administrative data

17 Heckman, Lalonde, and Smith (1999, page 1939) report, for example, that "econometric estimators that are valid for random samples can be invalid when applied to samples generated by matching procedures".

18 This assumes that the IV estimators in fact use a good instrumental variable. Use of a poor instrument could result in inconsistent estimates even if selection into the progam is based only on V1 and V2.

together with IV techniques to control for unobservable variables. That choice is reflected in the lower right hand corner of Table 1. The information in the table makes two important points about this analytical choice:

a. Properties of an approach that utilizes partial matching together with IV (Heckman) procedures are not completely understood. The basic problem is that the consistency of the IV procedures is based on the presumption that the comparison group is a random sample from some larger population of potential program participants. Partial matching prior to estimation of the "first stage" participation equation used in most IV procedures would not be consistent with that presumption. Determining how various IV procedures would perform with the partial matching that is possible in the EBSM context requires further research. In the absence of a clear finding that this approach yields consistent estimators in a particular approach, the panel would generally oppose such mixed approaches.

b. Other procedures — especially those that use difference-in-difference designs — may perform equally well in dealing with problems raised by unmeasured differences between participant and comparison groups in cases where those differences are constant over time. Hence, IV procedures are not necessarily the best strategy choice in all situations even when unobservable variables are believed to generate major differences between the participant and comparison groups.

Of course, these statements apply to just one potential analytical choice. Any others that might be suggested should be subjected to a similar analysis. It seems likely that the outcome from such an extended assessment of a variety of approaches would be that "it all depends on the nature of the unobservable variables". For this reason, the Panel believed that, to the extent feasible, a number of different analytical strategies should be explored in detail during the design phases of each evaluation and that *several* of the most promising approaches should be pursued in the actual analysis.

3. Precision of Estimates

Not only is it desirable that evaluation estimates of program impacts should be consistent, but, to be useful, they should also be relatively precise. In other words, the range of true impacts that is consistent with the estimates derived from an evaluation should be relatively narrow. If not, this range may be so broad that it is not possible to conclude anything about the effectiveness of a program — indeed an impact of zero (or even a negative impact) may be plausibly inferred from the data. Four general points should be made about achieving precision in the evaluations. First, achieving consistency is a goal that is logically prior to considerations of precision. If an evaluation offers no hope of providing consistent estimates it would do little good[19] to be assured that the (inaccurate) estimates that are provided are measured precisely. Second, achieving precision often becomes a matter of willingness to devote resources to an evaluation. In a sense, one can

[19] Actually, in some cases it may be possible to place bounds on the biases inherent in a particular estimation technique and one might opt for tradeoffs between precisely estimated effects with bounded biases and imprecisely estimated consistent estimates. Usually, however, the sources of inconsistency are not well enough understood to make such tradeoffs in any sort of comprehensive way.

always "buy" more precision by simply increasing sample sizes, so there will always be a tradeoff between precision and the willingness to pay for the information that an evaluation is expected to yield. A third point concerns the relationship between precision and the choice of estimators. Because some of the estimators recorded in Table 1 will require that estimates be based only on sub-samples of the data[20], this will involve an inescapable loss of precision. A more general related point is that for many of the estimators discussed in Table 1 it is not possible to make clear statements about precision. Indeed, in some cases, the uncertainty associated with an estimate cannot be calculated analytically and must be approximated using "bootstrap" simulation methods.

Given the complexities inherent in making statements about the precision of non-experimental estimates, it has become common to use illustrative precision estimates based on experimental designs in the planning of evaluations. Implicitly these illustrations assume the experimental calculations represent the best precision that might be expected in an evaluation and plans are based on this best-case scenario. Table 2 provides an example of this procedure. The calculations in the table show the size of impact that could be detected with 80 percent power assuming that a 95 percent one-tail test of significance is used in the evaluations. Standard deviations used in these calculations are drawn from the Nova Scotia feasibility study for the Medium Term Indicators (MTI) project. These may understate the actual precision that would be obtainable in the summative evaluations because they may have access to better control variables and statistical methodologies than were used in the Nova Scotia study. However, as we show in Section C, the precision estimates in the table are roughly consistent with what has been found in many actual experimental evaluations of active labour market programs.

TABLE 2
Detectable Differences with 80% Power and 95% Significance

Sample Sizes		Outcome (Assumed SD)		
Participant	Comparison	T4 Earnings (11,000)	Employed (0.370)	EI weeks (9.00)
2,000	1,000	1,103	0.037	0.90
2,000	2,000	901	0.030	0.74
3,000	1,500	901	0.030	0.74
4,000	2,000	780	0.026	0.64
1,000	1,000	1,274	0.043	1.04
1,000	400	1,685	0.057	1.38
500	400	1,911	0.064	1.56

20 This is especially true for the matching estimators that rely on survey data for which comparison group members who do not match participants would be dropped from the analysis.

Assuming that these figures are reasonably reflective of what can be expected from the evaluations, several observations might be made:

1. **Designs should take care to have adequate numbers of observations in the comparison groups.** Comparison of the second and third rows in the table, for example, shows that a balanced 2,000/2,000 sample yields exactly the same precision as an unbalanced 3,000/1,500 design. For the purposes of estimating differences between comparison and participant outcomes, the latter design wastes 500 interviews[21]. Importance of the size of the comparison group is also indicated by the final two rows in the table. These can be used to illustrate circumstances where matching using survey data causes drastic reductions in effective sample sizes. Clearly, in such situations, small program effects are unlikely to be detected at acceptable levels of significance and power.

2. **Evaluation designers should carefully examine whether it will indeed be possible to detect sufficiently small differences between comparison groups and participants in specific interventions.** The power calculations show that only very large effects will be detectable given the expected sample sizes for the analyses of specific Employment Benefits (EB) interventions. In combination with potential problems in selecting comparison groups for specific interventions, these findings suggest that expectations about measuring impacts of single interventions must be rather modest. Possibilities for estimating other sub-group or sub-provincial impacts must be similarly modest[22].

3. **Larger sample sizes should be considered for the evaluations.** Although the figures in the table indicate that most differences between participants and non-participants that might suggest cost-effectiveness should be detectable given the sample sizes envisioned (these are understood to be approximately 3,000 interviews), some of the figures, especially those for earnings gains, are problematic. The detectable differences could easily become unreasonably large if validity concerns led to deletion of large portions of some of the comparison groups. The extent to which these concerns would be mitigated with larger sample sizes should be examined in some detail in evaluation designs.

4. **Response rates should be carefully monitored in the evaluations.** Given the many issues that surround the validity of the analytical approaches in the evaluations, every effort should be made to ensure that the data do not fall prey to more mundane concerns (some of which are discussed in the next section). Availability of administrative data should aid in assessing the importance of non-response in the evaluations and an analysis of this issue should be expected as part of the evaluations.

---

21 Such unbalanced designs might be desirable if costs of participant and comparison cases differ substantially, but in most non-experimental evaluations that is not the case. Even in experimental evaluations it is usually the case that participant observations are more costly than control observations so that unbalanced designs should feature proportionally larger comparison groups.

22 The much larger samples available from the administrative data to be used in the MTI project suggest that this project is the suitable vehicle for examination of such sub-group impacts.

C. Sample Design

Three major questions must be faced in designing the actual samples to be used in the summative evaluations: (1) How is the participant group to be selected? (2) How are the administrative and survey data to be used in combination to select the comparison group(s)? and (3) How should resources be allocated to the survey samples[23]?

1. Participant Selection:

The Panel did not believe that there were major conceptual issues involved in selection of a participant sample for the evaluations (assuming that administrative data were of sufficiently high quality). Participants would be selected from administrative data in a way that best represents program activity during some period. The Panel did make three minor recommendations about this selection process:

a. Participants should be sampled over an entire year[24] so as to minimize potential seasonal influences on the results;

b. Participants should be stratified by Employment Benefit (EB) type and by location. This would ensure adequate representation of interventions with relatively small numbers of participants. It would also ensure representation of potential variations in program content across a region; and

c. The participant sample should be selected so that there would be a significant post-program period before surveying would be undertaken. Practically, this means that at least one year should have elapsed between the end of the sampling period and the start of the survey period[25].

2. Comparison Group Selection

Selection of comparison groups is one of the most crucial elements of the evaluation design effort. In order to simplify the discussion of this issue we assume that all analysis will be conducted using regression adjusted mean estimates of the impact of Employment Benefits and Support Measures (EBSM) on participants. That is, we, for the moment, disregard some of the estimation issues raised in Section B in order to focus explicitly on comparison group selection issues. Examining whether our conclusions here would be changed if different estimation strategies (such as difference-in-difference or IV procedures) were used is a topic requiring further research, as we discuss in Section F.

---

23 Administrative data are treated here as being plentiful and costless to collect. Sample sizes in the administrative data collection phase of the evaluations are therefore treated as unlimited. For specific, infrequent interventions this may not be the case, however, so we do briefly discuss sample allocations among interventions.

24 Use of a Fiscal Year would also facilitate comparisons to other administrative data — especially if start dates were used to define participation.

25 If surveys were conducted over an entire year this would permit two years to have elapsed since the program start date. If surveys were bunched so as to create interviewing efficiencies, the Panel recommended a longer period between the end of the sample period and the start of interviewing (perhaps 18 months or more).

Three data sources are potentially available for comparison group selection: (1) (Employment Insurance (EI)-related administrative data; (2) Canada Customs and Revenue Agency (CCRA) administrative data on earnings and other tax-related information; and (3) survey data. It is important to understand the tradeoffs involved in using these various data sources and how those tradeoffs might influence other aspects of the analysis.

a. **EI-related data:** These data are the most readily available for comparison group selection. Comparison group members could be selected on the basis of EI history and/or on basis of data on past job separations (the Record of Employment (ROE) data). Such matching would probably do a relatively poor job of actually matching participants' employment histories. That would be especially true for so-called "reachback" clients — those who are not currently on an active EI claim. Although it would be feasible to draw a comparison sample of individuals filing claims in the past, it seems likely that such individuals would have more recent employment experiences than would clients in the reachback group. Hence, even if matching solely on the EI data were considered suitable for the active claimant group (in itself a doubtful proposition) it would be necessary to adopt additional procedures for reachback clients.

b. **CCRA Earnings Data:** Availability of CCRA earnings data plays a potentially crucial role in the design of the summative evaluations. It is well known[26] that earnings patterns in the immediate pre-program period are an important predictor of program participation itself (Ashenfelter 1978, Heckman and Smith 1999). More generally, it is believed that adequately controlling for earnings patterns is one promising route to addressing evaluation problems raised by unobservable variables (Ashenfelter 1979). This supposition is supported by some early estimates of the impact of EBSM interventions in Nova Scotia (Nicholson 2000) which illustrate how CCRA data can be used in screening a broadly-defined comparison group to look more like a group of program participants[27]. Unfortunately, the extent to which these data will be available to EBSM evaluators is currently unknown. But, given the suggested sample selection and survey scheduling, it would have been feasible under previous data access standards to obtain an extensive pre-program profile for virtually all participants and potential comparison group members. Regardless of whether one opted for a general screening to produce a comparison group or used some form of pair-wise matching on an individual basis, it seems quite likely that a rather close matching on observed earnings could be obtained.

c. **Survey Data:** A third potential source of data for the development of a comparison sample is the follow-up survey that would be administered about two years after entry into the EBSM program. The advantage of this data source is that it provides

---

26 At least this is a common finding in evaluations of active labour market programs in the United States. Heckman, Lalonde, and Smith (1999) suggest that this "Ashenfelter dip" may be a worldwide phenomenon. The extent of the phenomenon among EBSM participants is unknown, although some preliminary work on data from Nova Scotia (Nicholson, 2000) suggests that the dip occurs in that program too.

27 Other variables available in CCRA tax data that may be used to match participant and comparison group members include, for example, total income, total family income, number of dependents, and so forth.

the opportunity to collect consistent and very detailed data from both participants and potential comparison group members about pre-program labour force activities and other information related to possible entry into EBSM interventions. These data can be used in a variety of methodologies (including both characteristic and propensity score matching and a variety of IV procedures) that seek to control for participant/comparison differences.

The potential advantages of using the survey data to structure analytical methodologies in the evaluations should not obscure the shortcomings of these data, however. These include:

- The survey data on pre-program activities will not be a true "baseline" survey. Rather, the data will be collected using questions that ask respondents to remember events several years in the past. **Errors in recall** on such surveys can be very large — and such errors will be directly incorporated into the methodologies that rely heavily on such retrospective survey data.

- Using the survey data to define comparison groups will necessarily result in some **reduction in sample sizes** ultimately available for analysis — simply because some of the surveyed individuals may prove to be inappropriate as comparison group members. The extent of this reduction will depend importantly on how much matching can be done with the administrative data. In the absence of the CCRA data such reductions could be very large. This would imply that a large amount of the funds spent on the survey might ultimately prove to have been expended for no analytical purpose.

- Finally, there is the possibility that reliance on the survey data to define comparison groups may **compromise the primary outcome data** to be collected by the survey. Most obviously this compromise would occur if the space needed in the survey to accommodate extensive pre-program information precluded the collection of more detailed post-program data. On a more subtle level, collecting both extensive pre- and post- program data in the same survey may encourage respondents to shade their responses in ways that impart unknown biases into the reported data. Although there is no strong empirical evidence on this possibility, survey staff will often check back-and-forth over a specific set of responses asking whether they are consistent. This may improve the quality of the data, but it may also impart biases to the reporting of retrospective data and therefore impact some estimates.

3. Suggested Approaches

This discussion suggests two approaches to the comparison group specification problem:

a. **The Ideal Approach:** It seems clear that, given the data potentially available to the evaluations, the ideal approach to comparison group selection would be to use both EI and CCRA data to devise a comparison sample that closely matched the

participant sample along dimensions observable in these two data sources. Pair-wise matching might be feasible with these large administrative data sets, but such an approach to matching may not be the best design from an overall perspective depending on the precise analytical strategies to be followed. For example, a strategy that focused more on sample screening to achieve matching may perform better in applications where there is an intention to utilize IV estimation procedures because screening may retain more of the underlying variance in variables that determine participation. Pair-wise matching also poses some logistical problems in implementation at the survey stage. Surveys could be conducted with **a random sample of pairs,** but non-response by one or the other member of a pair would exacerbate considerably the problem of non-response bias. Of course, no procedure of matching based on observable variables can promise that unobservable differences will not ultimately yield inconsistent results. Still, an approach that makes some use of EI and CCRA data to select the comparison sample would seem to be the most promising strategy in many circumstances. Some research on how first stage matching can best be utilized in an evaluation that will ultimately be based on survey data would clearly be desirable, however.

b. **The Alternative Approach:** If CCRA data are not available for sample selection in an evaluation, it would be necessary to adopt a series of clearly second-best procedures. These would start with some degree of rough matching using available EI data. The data sources and variables to be used might vary across the evaluations depending on the nature of local labour markets and of the individuals participating in the EB interventions. Following this rough matching, all further comparison group procedures would be based on the survey data. This approach would have three major consequences for the overall design of the evaluations:

   i. The **survey would have to be longer** so that adequate information of pre-program labour force activities could be gathered;

   ii. The **comparison group would have to be enlarged** relative to the "ideal" plan (see the next sub-section) to allow for the possibility of surveying individuals who ultimately prove to be non-comparable; and

   iii. The relative **importance of matching methods would have to be reduced** in the evaluations (if only because of the reduced sample sizes inherent in using a matching methodology based on the survey data) and the role for IV procedures[28] expanded. Because of this necessity of relying of IV procedures it may be necessary to devote additional resources to the development of "good" instruments either through additional data collection[29] or through some type of random assignment process.

---

28 Difference-in-difference methods might also be used more extensively though the use of such methods with data from a single survey opens the possibility of correlations in reporting errors over time biasing results.

29 For example, collecting data on geographical distances to service providers may provide a useful instrument in some locations.

4. Sample Sizes

The complexities and uncertainties involved in the design of the summative evaluations make it difficult to make definitive statements about desirable sample sizes. Still, the Panel believed that some conclusions about this issue could be drawn from other evaluations — especially those using random assignment. Because, in principle, randomly assigned samples pose no special analytical problems, they can be viewed as "base cases" against which non-experimental designs can be judged. Under ideal conditions where all the assumptions hold, a non-experimental design would be equivalent to a random assignment experiment once the appropriate analytical methodologies (such as pair-wise matching or IV techniques) have been applied. Hence, existing random assignment experiments provide an attractive model for the evaluations.

Table 3 records the sample sizes used for the analysis[30] of a few of the leading random assignment evaluations[31] in the United States:

TABLE 3
Analysis Sample Sizes in a Selection of Random Assignment Experiments

Evaluation	Experimental Sample Size	Control Sample Size	Number of Treatments
National JTPA	13,000	7,000	3
Illinois UI Bonus	4,186	3,963	1
NJ UI Bonus	7,200	2,400	3
PA UI Bonus	10,700	3,400	6
WA UI Bonus	12,000	3,000	6
WA Job Search	7,200	2,300	3
SC Claimant	4,500	1,500	3
Supported Work	3,200	3,400	1
S-D Income Maint.	2,400	1,700	7
National H.I.	2,600	1,900	3

Several patterns are apparent in this summary table:

- Sample sizes are all fairly large — control samples are at least 1,500 and more usually in the 2,000+ range;
- Single treatment experiments tend to opt for equal allocations of experimental and control cases[32];

30 These "final" sample sizes allow for survey and item nonresponse. Initial sample sizes would have to be increased to allow for such attritions.

31 For a summary of many of these evaluations together with an extensive set of references, see Greenberg and Shroder (1997).

32 Such an allocation would minimize the variance of an estimated treatment effect for a given evaluation budget assuming that treatment and control cases are equally costly.

- Evaluations with multiple treatments allocate relatively larger portions of their samples to treatment categories. Usually the control groups are larger than any single treatment cell, however; and

- Although it is not apparent in the table, many of the evaluations utilized a "tiered" treatment design in which more complex treatments were created by adding components to simple treatments (this was the case for many of the UI-related evaluations, for example). In this case, the simple treatments can act as "controls" for the more complex ones by allowing measurement of the incremental effects of the added treatments[33]. Hence, the effective number of "controls" may be understated in the table for these evaluations.

Because many of these evaluations were seeking to measure outcomes quite similar to those expected to be measured in the EBSM evaluations, these sample sizes would appear to have some relevance to the EBSM case. Specifically, these experiences would seem to suggest effective comparison sample sizes of at least 2,000 individuals[34]. The case for such a relatively large comparison sample is buttressed by consideration of the nature of the treatments to be examined in the EBSM evaluation. Because the five major interventions offered under regional EBSM programs are quite different from each other, it will not be possible to obtain the efficiencies that arise from the tiered designs characteristic of the UI experiments[35]. Heterogeneity in the characteristics of participants in the individual EBSM interventions poses an added reason for a large comparison group. In evaluating individual interventions it is likely that only a small portion of the overall comparison group can be used for each intervention. Again, the case for a large comparison sample is a strong one.

Interest in evaluating the specific EB interventions raises additional complications for sample allocation. In most regions SD interventions are by far the most common. A simple random sample of interventions would therefore run the danger of providing very small sample sizes for the other interventions. Hence, it seems likely that most evaluations will stratify their participant samples by intervention and over-weight the smaller interventions[36]. If an evaluation adopts this structure, deciding on a comparison structure is a complex issue. Should the comparison sample also be stratified so that each sub-group mirrors individuals in the intervention strata? Or should the comparison sample seek to match the characteristics of a random sample of participants? The latter approach might

---

33 In many of the evaluations, however, the less elaborate treatments often prove to be the most effective. That is the case in practically all of the UI-related experiments.

34 This conclusion is roughly consistent with the illustrative power calculations presented in Section B (which are based on variances observed in the Nova Scotia data). It should again be noted that the sample sizes suggested here are effective sample sizes. In other words, these would be the number of cases available for analysis, net of out-of-scope cases, non-responses to surveys, and so forth. In addition, where income data are required, individual projects may be advised to implement the surveys in the field at approximately tax time so that respondents will have all their tax data readily available and fresh in their minds during the interview.

35 A possible tiered design would be to adopt an EAS-only cell in some of the evaluations, however. Experiences from the UI experiments in the United States suggests that the EAS-only treatment might indeed have some detectable effects.

36 More generally, the participant sample could be allocated to specific interventions on the basis of regional policy interest in each such intervention. This might dictate an allocation plan that represented a compromise between proportionate representation and stratification into equal size cells.

be preferable if IV procedures were intended to play a major role in the analysis because such a broader comparison group might aid in predicting assignment to treatments. On the other hand, the use of intervention-specific comparison groups might be preferable for use with simpler estimation procedures. Evaluators should be expected to address such matters carefully in their designs.

5. Issues in Implementation

Implementation of any sort of matching strategy in an evaluation that relies on surveys poses special problems. The general goal is to avoid "wasting" interviews of comparison group members that will never be used in the subsequent analysis. This usually requires that some significant portion of interviews of the participant sample be completed before the comparison interviews are fielded. In that way, the characteristics of the comparison sample can be adjusted based on the experiences encountered with the participant sample[37]. An even more complex surveying strategy might be necessary if only minimal administrative data are available for the sample selection process. In this case, it may be necessary to include some form of screening questions in the surveys of comparison cases so that surveys of non-comparable cases can be cut short before large amounts of time have been expended. This procedure has been suggested, for example, for dealing with the problem of finding comparison cases for participants in the reachback group. Because it is possible to build in biases through such survey-based sample selection procedures, they should be adopted only with extreme caution.

37 The actual tailoring of such procedures to deal with survey nonresponse can be a very tricky issue, especially if participant and comparison groups have differential rates of nonresponse. This is another reason why issues related to nonresponse warrant a prominent place in evaluation designs.

D. Outcome Specification and Measurement

Four criteria guided the Panel's thoughts on outcome specification and measurement: (1) It seems likely that for most evaluations some of the key measures[38] will focus on employment in the post-program[39] period; (2) Given this focus, it seems clear that sufficient employment information should be collected so that a variety of specific measures can be calculated. This will aid in the tailoring of outcome measures to provide an accurate reflection of specific program purposes[40]; (3) Data on a number of other key socio-economic variables should be collected — primarily for use as control variables in studying employment outcomes; and (4) Survey data collection techniques should strive for direct comparability across different regional evaluations.

Seven specific actions that might serve to meet these criteria are:

1. **Barring special considerations, the Follow-up survey should occur at least two years after Action Plan start dates.** The intent of this recommendation on timing is to offer a high probability that interventions are completed well before the interview date. Because the first evaluations are contemplating Fall 2001 interview dates, this would require that action plans with start dates during FY99 (April 1, 1998 — March 31, 1999) be used. Evaluations with later interview dates might focus on FY00 instead.

2. **When possible, similar employment history questions should be used in all of the evaluations.** Because post-start and post-program employment will probably be the focus of most of the EBSM evaluations, it seems clear that significant survey resources should be devoted to its measurement. Prior studies have documented that use of different data collection instruments can lead to quite different estimated results (Heckman, Lalonde, and Smith 1999). To avoid this source of variation, evaluators should be encouraged to use the same question batteries. Similarly, data cleaning and variable construction routines should be coordinated across the evaluations. Evaluators should also consider how the availability of documentation in the possession of respondents (for example, tax documents) can be best utilized in the collection of these data.

---

38 Other measures might include program completion, Employment Insurance (EI) collections, and subsequent enrollment in additional Employment Benefits and Support Measures (EBSM) interventions.

39 In actuality, employment should usually be measured from the entry date into the program (and from a similar date for comparison cases) because this permits the most accurate appraisal of foregone employment opportunities caused by program participation.

40 In Section F we discuss the need to develop a research program to study how outcomes should be related to EBSM and more general societal program goals.

3. **A number of employment-related measures should be constructed.** The goal of this recommendation is to ensure that the outcome measures being used in the evaluations are in fact appropriate to the intended program goals. Although all evaluations would be expected to construct the same basic measures (such as weeks employed during the past year, total earnings during that period, number of jobs held, and so forth), there would also be some flexibility for specific regions to focus on the measures they considered to be most appropriate to the package of benefits being offered. For example, outcomes for clients in Skills Development interventions might focus on wage increases in the post program period or on changes in the types of jobs held. Outcomes for participants in Targeted Wage Subsidy programs might focus on successes in transitioning to unsubsidized employment. And it may prove very difficult to design ways of measuring the long-term viability of the self-employment options pursued by some clients. As mentioned previously, there is a clear need for further research on precisely how measured outcomes and interventions will be linked. On the more conceptual level there is the need to show explicitly how the outcomes that are to be measured in the evaluations are tied to the general societal goals of the (EBSM) program (as stated, for example, in its enabling legislation).

4. **It is desirable that a core module for collecting data on other socio-economic variables be developed. Such a module could be used across most of the evaluations.** The goal of such a module would be to foster some agreement about which intervening variables should be measured and to ensure that these would be available in all of the evaluations. In the absence of such an agreement it may be very difficult to compare analytical results across regions because the analyses from different evaluations would be controlling for different factors. Pooling of data for cross-region analysis would also be inhibited if such a module were not used. Clearly Human Resources Development Canada (HRDC) has a direct interest in such cross-region analyses both because they may be able to make some estimates more precise and because they may be able to identify important determinants of differential success across regions[41]. Hence, it should consider ways in which all evaluators could be encouraged to use similar core modules — perhaps by developing them under separate contract.

5. **Additional follow-up interviews should be considered, at least in some locations.** Although most evaluations will probably utilize a one-shot survey approach, the Panel believed that evaluators should be encouraged to appraise what might be learned from a subsequent follow-up (perhaps 24 months after the initial survey). It seems likely that such additional data collection would be especially warranted in cases for which interventions promise only relatively long term payoffs. It seems likely that additional follow-up interviews, if they were deemed crucial to an evaluation, would be independently contracted. Regardless of whether a follow-up interview is included as part of an evaluation design, the Panel believed

41 The desirability of pooling data from different evaluations is discussed further in Section F.

that HRDC should make arrangements that would enable evaluation participants to be followed over time using administrative data on EI participation and (ideally) earnings (see the next point).

6. **Administrative data should be employed to measure some outcomes in all of the evaluations.** Timing factors may prevent the use of administrative earnings data (from T-4's) to measure outcomes in the evaluations as currently contracted (though this could be part of a follow-up contract), but EI administrative data should be utilized to the fullest extent practicable. These data can provide the most accurate measures of EI receipt and can also shed some light on the validity of the survey results on employment. Administrative data can also be used in the evaluations to construct measures similar to those to be constructed in the medium term indicators (MTI) pilot project thereby facilitating comparisons between the two studies (see Section E below). Using administrative data to measure outcomes also has benefits that would extend far beyond individual evaluation contracts. In principle it should be possible to follow members of the participant and comparison groups for many years using such data. Use of these data would aid in determining whether program impacts observed in the evaluations persisted or were subject to rapid decay. It is also possible that assembling the longer longitudinal data sets made possible by using administrative data could shed some light on the validity of the original impact estimates by making fuller use of measured variations in the time series properties of earnings for participant and comparison groups.

7. **Cost-benefit and cost-effectiveness analyses should be considered, but they are likely to play a secondary role in the evaluations.** Estimates of impacts derived in the evaluations could play important roles in providing the bases for cost-benefit and cost-effectiveness analyses. Development of a relatively simple cost-effectiveness analysis would be straightforward assuming data on incremental intervention costs are available. The utility of such an analysis depends importantly on the ability to estimate impacts of specific interventions accurately — both from the perspective of quasi-experimental designs and the adequacy of sample sizes to provide sufficiently detailed estimates. Still, it may be possible to make some rough cross-interventions comparisons.

   Conducting extensive cost-benefit analyses under the evaluations might present more significant difficulties, however, especially given the sizes of budgets anticipated. Some of the primary obstacles to conducting a comprehensive cost-benefit analysis include the possibility that many of the social benefits of the EBSM program may be difficult to measure, that estimating the long-run impacts of the programs may be difficult and that the overall size of the program may make displacement effects significant in some locations. Methodologies for addressing this latter issue in any simple way are especially problematic. For all of these reasons, the panel believed that the planned modest budgets of the various evaluations would not support the kind of research effort that would be required to mount a viable cost-benefit analysis. However, the panel strongly believes that some sort of cost-benefit analysis should be the ultimate goal of the evaluations

because stakeholders will want to know whether the programs are "worth" what they cost. Hence, it believes that it may be prudent for HRDC to consider how a broader cost-benefit approach might be conducted by using the combined data[42] from several of the evaluations taken together as part of a separate research effort.

[42] At a minimum, contractors should be encouraged to provide public use data sets that might be combined by other researchers in the development of cost-benefit or other types of analysis.

E. Coordination with the medium term indicators (MTI) Project

The Labour Market Development Agreements (LMDA) evaluations will play an important role in the process leading to the development of medium term indicators for the Employment Benefits and Support Measures (EBSM) program (the "MTI project"). Because these indicators will be constructed mainly from administrative data, the evaluations offer the opportunity to appraise the potential shortcomings associated with indicators that are based on relatively limited data only. Such a comparison can help to clarify how impacts measured in the MTI project may fail to capture all of the effects (both positive and negative) of Employment Benefit (EB) interventions. The availability of a richer data set in the evaluations may also suggest simple ways in which planned medium term indicators might be improved (say by combining data from several administrative sources or by using innovative ways of aggregating over time). Alternatively, the MTI project will, by virtue of its much larger sample sizes and on-going operations, permit the study of effects on sub-groups and recent program innovations that cannot be addressed in the LMDA evaluations. MTI-generated impacts from such investigations will be more meaningful if the evaluations have established baseline validity measures that show how such outcomes relate to the more detailed impact estimates from the LMDA evaluations. For all of these reasons, coordination between the two projects is essential. In order to achieve that coordination the panel suggested the following general guidelines:

1. **Evaluation samples should be drawn in a way that facilitates comparison with MTI results.** Many of the suggestions in sections A and B above are intended to achieve this goal. In general, it would be hoped that the evaluation samples could be regarded as random samples of the larger populations that would be used for MTI construction during a given period. In this way, the evaluation results can be viewed as providing direct assessments of the MTI measures. Consultations during the design phases of the LMDA evaluations will be required in order to ensure that this goal is achieved.

2. **LMDA evaluation contractors should, where feasible, develop MTI-like measures for their research samples (or make it possible for other researchers to do so).** Examining the performance of the medium term indicators and how they might be improved will require the use of micro-data from the evaluations. That will not be possible unless some care is taken to ensure that the appropriate administrative data has been added to the research files. Of course, some data to be used to construct MTI's may not be available to the evaluators in a timely manner (that will probably be the case for more recent Canada Customs and Revenue Agency (CCRA) data). In such cases, research files should be developed in ways that would permit these data to be added at a later date and used to augment the analysis then.

3. **Efforts should be made to coordinate scheduled evaluations with the MTI development process.** Many of the coordination possibilities between the LMDA evaluations and the MTI project will be lost or impaired if these two efforts are not conducted on roughly the same time frame. Because the scheduling of the evaluations is more-or-less determined under the LMDAs, policy-makers interested in the development of MTI's should, when feasible, try to match this schedule.

F. Summary of Issues Requiring Further Consideration at Implementation

This paper presents a state of the art generic framework for the evaluations of the Employment Benefits and Support Measures (EBSMs). In applying the methodologies outlined in this paper, those implementing the evaluations should take into consideration conditions unique to their particular jurisdictions. In the process, any assumptions set against the framework as well as modifications and extensions should be defensible. They should be well documented and rationalized on the basis of the evidence, the justification for the application of the techniques, and the underlying knowledge and theory. The credibility of results can benefit from the simultaneous use of several techniques to obtain multiple lines of evidence. Moreover, to reflect some of the uncertainties inherent in quasi-experimental designs, the implementers should be encouraged to provide ranges of estimates, as opposed to single point estimates.

This paper has provided a very general treatment of some of the major issues that can be expected to arise as the designs of the EBSM evaluations proceed. The intent here was not to provide an explicit blueprint for each such evaluation, but rather to describe some of the universal, methodological questions that will have to be addressed. The Panel believes that the EBSM evaluations offer the possibility of providing cutting-edge evidence in evaluation research while, at the same time, providing substantial information that will be both valid and useful to policy-makers.

The panel identified a number of areas that require further attention. It therefore falls to the implementers to fill in the details.

1. **Division of labour among evaluations should be beneficial.** Because not all of the evaluations can be expected to do "everything", the Panel believes that some division of labour among them would be appropriate and beneficial to obtaining the best overall results. Three examples are: (1) A more in-depth study of some interventions (e.g. self-employment) in some evaluations than in others; (2) a greater focus on difficult to measure non-employment outcomes in some evaluations than in others; or (3) any opportunity in an evaluation to implement random assignment should be embraced.

2. **Procedures for appraising the results of the evaluations and for the reporting of those results should be developed.** Although individuals not in the research community may be expecting to know *the results* of the EBSM evaluations, it seems much more likely that the actual results will be quite varied across regions. Some of that variation may arise from true differences in program impacts, some from differences in program design and implementation, some from differences in evaluation design, some from differences in labour demand across the regions, and

some from purely random factors (such as labour market differences). The Panel believes that a haphazard reporting of these variations may be detrimental to the overall perceived validity of the evaluation effort. In order to mitigate such problems, it would seem desirable at the outset to establish some sort of reporting procedures for the evaluations that would place the results in their overall context. This context would articulate the unique aspects of each individual evaluation, along with a view of how its results fit into an overall pattern of interpretation.

3. **Ways should be considered in which the results from the evaluations can be pooled in subsequent analyses.** The Panel believes that the pooling of results from across the evaluations, although it may pose problems of statistical non-comparability, should be an essential part of the EBSM evaluation process and contractors should be encouraged to make their data available for such purposes. Such pooling would have a number of advantages including:

    - Added sample sizes so that precision of estimates can be improved;
    - The ability to study the effects of potential differences in program content;
    - Insights on how the Employment Benefit (EB) interventions perform in different labour markets;
    - Provide greater possibilities for examining outcomes for subgroups of participants in the EBSM program; and
    - Open the possibility for exploring issues that may be beyond the scope of any single evaluation — such as undertaking a detailed cost-benefit analysis of the overall program.

To achieve these important goals, it will be necessary to impose some degree of comparability on data collection and accessibility across the evaluations. Such standards must be built in from the start — it may prove impossible to pool the data across evaluations if each is permitted to have complete freedom in defining unique approaches to data issues.

4. **To the extent possible, common data collection instruments across the evaluations should be used.** This conclusion derives from the three prior considerations. All issues of reporting and pooling of results from the evaluations will be made more difficult if each is based on unique data collection instruments. Hence, the Panel strongly believes that some common modules (especially those related to labour market histories and outcomes) should be included across the evaluations. Developing such modules will require significant consultation among various stakeholders in the EBSM program both to ensure that the data collection instrument is of a high caliber and to ensure that information is collected on all policy-relevant outcomes (including a variety of potential socio-economic outcomes that have not been explicitly examined in this paper).

5. **How much matching should occur before selection of the survey samples?** As discussed in Sections B and C, it is not possible to design a survey sampling plan for the evaluations until the extent of matching that can be accomplished with administrative data is clarified. For this reason, the panel believed that it is important to undertake some early work with the administrative data to determine the limitations of the matching process. In general, these limitations are expected to be more severe if Canada Customs and Revenue Agency (CCRA) data are not available for sample selection in the earliest evaluations. In the application of the methodology, it is essential that the shortcomings of using only Employment Insurance (EI) data to select comparison groups be documented. Means for addressing these shortcomings should also be addressed in the design and implementation of the survey. Especially important is to determine how the allocation of sample between participant and comparison cases will be influenced by the inability to identify good comparison matches with the EI data alone.

6. **What role should IV estimates play in the evaluations?** Implementation of IV estimators raises some of the most difficult issues in the evaluations. This is the case both because the identification of situations in which such estimators would yield consistent treatment effect estimates is often ambiguous and because assessing the estimators, once obtained, is as much art as science. For these reasons, the panel strongly believes that role of IV estimators in the evaluations should be examined. Some of the issues about utilizing such estimators that should be addressed are:

    - What variables will be used to achieve the identification the IV estimators require? How will these data be collected? How should identifying restrictions be tested?

    - How will the properties of IV estimators be affected by the ways in which the participant and comparison samples are selected? Can simulation analysis contribute to an understanding of the relationship between IV estimation and sample matching?

    - How can the robustness of IV estimators be assessed? How should potential alternative specifications for IV estimators be reported and evaluated?

7. **Further consideration of sample size is required for each particular application.** The calculations presented in this paper suggest that sample sizes in the range of 2,000 participants and 2,000 comparison cases are what might be minimally needed to ensure that estimates from the evaluations will be sufficiently precise so that it will be possible to detect policy relevant impacts. But this assessment of precision was made on the basis of very limited information or solely from comparisons to other studies. Clearly, a review of this issue, as it applies to each specific evaluation, is warranted before major resources are committed. This review should explore such issues as: (1) The desire to estimate impacts for specific interventions; (2) How (if at all) the comparison group can be "shared" among differing interventions; (3) How analysis methods are likely to affect available sample sizes; and (4) How the likelihood of non-response affects sample size computations.

References

Ashenfelter, Orley. "Estimating the Effects of Training Programs on Earnings" *Review of Economics and Statistics* January, 1978, pp. 47-57.

Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings with Longitudinal Data" in F. Bloch (ed.) *Evaluating Manpower Training Programs.* Greenwich (CT), JAI Press, 1979, pp. 97-117.

Ashenfelter, Orley and Card, David "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs" *Review of Economics and Statistics*, June, 1985, pp. 648-660.

Dehejia, Rajeev and Wahba, Sadek. "Propensity Score Matching Methods for Nonexperimental Causal Studies." National Bureau of Economic Research (Cambridge, MA) Working Paper NO. 6829, 1998.

Dehejia, Rajeev and Wahba, Sadek. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, December 1999, 94(448), pp. 1053-62.

Greenberg, David and Shroder, Mark. *The Digest of Social Experiments, Second Edition.* Washington, D.C. The Urban Institute Press, 1997.

Heckman, James. "Varieties of Selection Bias." *American Economic Review Papers and Proceedings,* May 1990, 80(2), pp. 313-18.

Heckman, James and Hotz, Joseph. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association,* December 1989, 84(408), pp. 862-74.

Heckman, James, Ichimura, Hidehiko; Smith, Jeffrey and Todd, Petra. "Characterizing Selection Bias Using Experimental Data." *Econometrica,* September 1998a, 66(5), pp. 1017-98.

Heckman James, Ichimura, Hidehiko and Todd, Petra. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies*, October 1997, 64(4), pp. 605-54.

Heckman, James, Ichimura, Hidehiko and Todd, Petra. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies*, April 1998b, 65(2), pp. 261-94.

Heckman, James, Lalonde, Robert and Smith, Jeffrey. "The Economics and Econometrics of Active Labor Market Programs," in Orley Ashenfelter and David Card, eds., *Handbook of labor economics,* Vol. 3A. Amsterdam: North-Holland, 1999, pp. 1865-2097.

Heckman, James and Smith, Jeffrey. "The Preprogramme Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies." *Economic Journal*, July 1999, 109(457), pp. 313-48.

Ichimura, Hidehiko and Taber, Christopher. "Direct Estimation of Policy Impacts." National Bureau of Economic Research (Cambridge, MA) Technical Working Paper No. 254, 2000.

Imbens, Guido and Angrist, Joshua. "Identification and Estimation of Local Av erage Treatment Effects." *Econometrica*, March 1994, 62(2), pp. 467-75.

LaLonde, Robert. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, September 1986, 76(4), pp. 604-20.

Newey, Whitney and Powell, James. "Instrumental Variables for Nonparametric Models." Unpublished manuscript, Princeton University, 1989

Nicholson, Walter. "Assessing the Feasibility of Measuring Medium Term Net Impacts of the EBSM Program in Nova Scotia" Working Paper prepared for HRDC, March, 2000.

Nicholson, Walter. "Design Issues in the Summative Evaluations: A Summary of Meetings..." HRDC, March, 2001.

Rosenbaum, Paul and Rubin, Donald. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, April 1983, 70(1), pp. 41-55.

Smith, Jeffrey and Todd, Petra. "Reconciling Conflicting Evidence on Performance of Propensity Score Matching Methods." *American Economic Review Papers and Proceedings*, May 2001, 91(2), pp. 112-18.

Annex A[43]: Summative Evaluations of Employment Benefits and Support Measures and Labour Market Development Agreements — Possible Program Effects to be Examined/Measured

This annex identifies a range of possible program effects for the evaluation. This "possible program effects approach" is consistent with the framework used by the Office of the Auditor General in the recent past to audit program effectiveness measurement by federal departments and agencies.

The intention for evaluation purposes is *not* to measure individual results versus some desired quantitative goal or detailed quantitative benchmark which is expected to be achieved. Rather, the approach is to examine and measure, as comprehensively as possible, what it is that the EBSM/LMDA-funded programs have been doing (their activities) and what the results have been.

The intention is to undertake the same basic evaluation approach across all of the regions and then develop an aggregate analysis at the national level with respect to overall EBSM/LMDA program results. Provision is also to be built-in for local flexibility in terms of subject matter areas covered by providing for any further, additional evaluation questions/indicators which individual regions may wish to include.

It is expected that, once the evaluation process proceeds, consultants selected will prepare methodology reports that will deal in more detail with the measurement process in the areas identified.

---

43 This Annex does not form part of the expert panel's framework. Their deliberations were devoted mostly to the methodology outlined in the body of the report. Discussions of specifically what to measure were very limited. The clarification of measurement issues may provide an important task for the panel in the future.

The possible effects listed below in the section entitled "Possible Program Effects..."are based on issues that have considered important in many employment-related policies and studies implemented by various researchers and levels of government nationally and internationally. In summary, important sources for issues with respect to EBSM evaluations can be derived from:

- the objectives of the LMDAs/EBSM as stated in the Employment Insurance (EI) legislation and labour market agreements;
- issues raised in the context of evaluations studies, policy research, policy statements, etc., as they relate to employment;
- issues studied and raised in the formative evaluations of the LMDAs/EBSM;
- issues raised in the early planning of summative evaluations and the Medium-term Indicators pilot project;
- possible effects that investments of this nature may have in the labour market and the economy broadly speaking; and
- developing social and economic trends in Canada and among its trading partners.

As indicated in the body of this report, the methodology includes several measurement techniques that may be applied to a broad range of outcomes/indicators. Where appropriate, therefore, Human Resources Development Canada (HRDC) expects that outcomes will be measured as follows: post- program outcomes, differences between pre and post program experiences, and comparisons between participant and non-participant experiences.

POSSIBLE PROGRAM EFFECTS
PROGRAM OUTCOMES

INDIVIDUAL CLIENTS ASSISTED

- PERCENT OF PARTICIPANTS WHO FOUND JOBS IN POST-PROGRAM PERIOD
- PERCENT WEEKS WORKED OVER POST-PROGRAM PERIOD
- AVERAGE DURATION OF EMPLOYMENT SPELLS
- AVERAGE LEVELS OF EI/SOCIAL ASSISTANCE COLLECTED
- PERCENT OF WEEKS ON EI AND/OR /SOCIAL ASSISTANCE IN THE POST PROGRAM PERIOD (worker self-sufficiency — reduced reliance on EI and social assistance)
- PRE- AND POST-PROGRAM UTILIZATION OF EI vs SA.
- AVERAGE EARNINGS VS LOW-INCOME CUT-OFF (or other market- or community-based measures of earnings/income)

- AVERAGE INCOME VS LOW-INCOME CUT-OFF (or other market- or community-based measures of earnings/income)

- QUALITY OF LIFE INDICATORS (FOCUS ON IMPROVEMENTS FOR INDIVIDUALS VS NO CHANGE: — MOTIVATION; SATISFACTION WITH LIFE). (N.B. This aspect is included as part of the proposed evaluation approach (a) because it is a stated overall goal for the department; and (b) because the general technical literature (including material from other countries) identifies that such effects can be associated with LMDA-type program activities).

- GEOGRAPHIC MOBILITY/RELOCATION OF WORKER CLIENTS

KEY CLIENTELE CHARACTERISTICS (Focus on key types of problems being encountered by individual clients assisted under the LMDA arrangements).

- AGE/SEX VS TOTAL UNEMPLOYED (YOUTH UNEMPLOYMENT / GENDER EQUITY ISSUES)

- PROPORTION OF WORKER CLIENTS ENCOUNTERING CHRONIC EMPLOYMENT DIFFICULTIES/EMPLOYMENT BARRIERS PRIOR TO PROGRAM ASSISTANCE (To be based on a literature search on prevailing types of employment barriers)

- PROPORTION OF WORKERS CLIENTS IDENTIFIED AS DISPLACED WORKERS PRIOR TO PROGRAM ASSISTANCE

- EDUCATIONAL LEVELS (FUNCTIONAL LITERACY ISSUES)

- PRE- AND POST- OCCUPATIONAL CATEGORIES AND SKILL LEVELS (Skill deficiency/skill development)

- APPRENTICES (DESCRIPTIVE MATERIAL ONLY)

LOCAL COMMUNITIES

- LABOUR MARKET DEVELOPMENT TO ASSIST COMMUNITIES (E.G. ASSIST LOCAL ECONOMIC DEVELOPMENT AGENCIES/ LOCATION OF INDUSTRY) (DESCRIPTIVE MATERIAL ONLY)

GOVERNMENTS

- COST-EFFECTIVENESS OF PROGRAM INTERVENTIONS (E.G. Net improvements in employability and employment duration within various categories of clients vs the costs of the program assistance provided).

- EBSM/LMDA'S AS AN OCCUPATIONAL SUPPLY CHANNEL (Assessment of scale of LMDA activities vs labour expansion and replacement needs of the economy).

- GROSS SAVINGS TO THE EI ACCOUNT AS A RESULT OF EBSM/LMDA

- NET SAVINGS TO THE EI ACCOUNT AS A RESULT OF THE EBSM/LMDA

FOLLOW-UP TO FORMATIVE STUDIES

- SELECTED ISSUES RAISED IN THE FORMATIVE STUDIES

- SELECTED ISSUES THAT HAVE EMERGED SINCE THE COMPLETION OF THE FORMATIVE STUDIES

La conception des évaluations sommatives pour les prestations d'emploi et mesures de soutien (PEMS)

(à l'intention du groupe d'experts de Développement des ressources humaines Canada (DRHC) sur la conception des évaluations sommatives pour le programme des PEMS)

Rédigé par :
Walter Nicholson

septembre 2001

SP-AH168-09-01F
(also available in English)

Table des matières

Préface i

A. Définir les « participants » aux fins des évaluations des prestations d'emploi et mesures de soutien (PEMS) 1

B. Méthodes d'analyse à utiliser dans les évaluations 5

C. Plan d'échantillonnage 17

D. Spécification et mesure des résultats obtenus par les participants 25

E. Coordination avec le projet des indicateurs à moyen terme (IMT) 29

F. Sommaire des questions sur lesquelles il faudra revenir au moment de la mise en œuvre 31

Annexe A : Évaluations sommatives des prestations d'emploi et mesures de soutien ainsi que des ententes sur le développement du marché du travail — répercussions éventuelles des programmes à examiner/mesurer 37

Préface

Développement des ressources humaines Canada (DRHC) et ses partenaires provinciaux et territoriaux reconnaissent l'importance des constats des évaluations lorsqu'il s'agit d'examiner les répercussions des prestations d'emploi et mesures de soutien ainsi que des ententes sur le développement du marché du travail (PEMS/EDMT). On trouve des dispositions concernant l'évaluation de ces éléments autant dans la législation sur l'assurance-emploi que dans les EDMT.

En général, l'évaluation des interventions en matière d'emploi constitue une tâche complexe. Dans le cas des évaluations sommatives des PEMS/EDMT, la grande diversité des collectivités et des modèles de mise en œuvre à l'échelle du pays, le nombre d'évaluations distinctes à réaliser et le calendrier d'exécution de chacune ajoutent à la complexité de la tâche.

Les évaluations appellent une certaine réflexion créative jumelée à une connaissance approfondie des techniques de mesure, notamment parce qu'elles doivent combiner la connaissance de théories compliquées du marché du travail et de l'économie, de modèles sophistiqués de mise en œuvre de programmes et d'acquisition du savoir, d'éventails complexes de répercussions éventuelles prévues et imprévues et de modèles statistiques fort structurés qui serviront à estimer, à partir de listes partielles de concepts, des résultats et des répercussions qui, bien souvent, ne peuvent être mesurés directement.

Par leur nature même, les évaluations des PEMS/EDMT nécessitent l'appel à de nombreux partenaires ayant divers intérêts, y compris les décideurs, les gestionnaires de programme, les agents de mise en œuvre, les consultants et les membres du public (comme les participants à un programme et les contribuables). Les membres de ces groupes possèdent des niveaux différents de connaissance des principes, des normes et des techniques d'évaluation ainsi que pour l'interprétation des constats. Pour que l'évaluation soit efficace, il est essentiel d'établir un niveau élevé de confiance entre tous les partenaires et toutes les parties intéressées. Il est donc important que la méthodologie utilisée soit fort justifiable et qu'elle soit présentée de façon claire et tout à fait transparente.

DRHC a confié à un groupe d'experts le mandat d'élaborer un cadre générique objectif et inspiré de l'état actuel des connaissances pour l'évaluation des PEMS/EDMT. Les experts ont élaboré le cadre de façon à offrir un fondement méthodologique commun et solide pour les diverses évaluations. Selon un des principes directeurs, il devait comporter assez de souplesse pour satisfaire aux exigences en matière de compte rendu aux niveaux national, provincial, territorial et local. Une de ses principales caractéristiques est qu'il est détaillé et qu'il peut être adapté facilement à des situations opérationnelles et à des contenus bien précis. La méthodologie proposée est fondée sur les plus récentes approches théoriques et pratiques pour l'évaluation ainsi que sur les points de vue et les expériences des experts en ce qui a trait à son application.

Le cadre proposé par le groupe d'experts est résumé dans le présent rapport. Cela devrait faciliter la conclusion d'une entente en ce qui concerne l'adoption d'une méthodologie, de sorte qu'au moment de communiquer les résultats, on pourra se concentrer sur l'interprétation des constats concernant les répercussions probables des PEMS/EDMT. Le rapport porte presque exclusivement sur les méthodes qui permettront de mesurer les résultats, plutôt que sur les éléments à mesurer. Le cadre d'évaluation proposé peut s'appliquer à un vaste éventail de résultats (notamment à ceux dont il est question à l'annexe A). DRHC tient à remercier M. Walter Nicholson, Jeff Smith, Chris O'Leary, Paul Decker et Joe Michalski de leur excellent travail.

DRHC présente ce rapport par souci de transparence en vue de faire progresser notre connaissance de la méthodologie des évaluations et faciliter les discussions à son sujet ainsi que pour assurer les meilleures évaluations sommatives possibles des PEMS.

Le présent rapport expose un cadre général qui peut constituer la base de toute évaluation bien conçue des prestations d'empli et mesures de soutien (PEMS)[1]. Il détermine les éléments qui devraient être considérés, passe en revue les points forts et les points faibles de diverses formules envisageables pour établir des mesures et souligne les questions qui pourraient faire l'objet de recherches supplémentaires[2].

Même s'il examine certaines questions de manière approfondie, ce document ne se veut pas le plan détaillé et définitif d'une évaluation en particulier. Comme on pourra le voir, le cadre général devra être adapté (modifié et élargi) afin de rendre compte de la singularité des exigences propres à chaque évaluation. Par exemple, tout plan d'évaluation définitif se fondera sur les buts de l'évaluation, la nature du programme à examiner, la qualité des données disponibles, de même que des considérations pratiques comme le calendrier de réalisation et les ressources disponibles.

La diffusion des résultats des évaluations et leur validité perçue représentent des éléments importants à prendre en considération dans l'évaluation des PEMS. Même si ce document tente de faire le point sur l'état actuel des connaissances relatives aux méthodes d'évaluation, il ne prétend pas définir ce qu'est une « bonne » évaluation, ni la façon dont les résultats obtenus par de telles méthodes devraient être utilisés. Pour être en mesure de juger de la validité globale des éléments de preuve, il faut particulièrement : 1) comprendre *a priori* l'applicabilité et les limites de l'approche générale utilisée (par exemple, les limites des modèles quasi expérimentaux); et 2) examiner a posteriori la qualité des données recueillies et la robustesse des méthodes d'estimation employées. C'est particulièrement important dans le contexte des évaluations des PEMS, car l'un des grands objectifs de la plupart de celles-ci sera probablement de recueillir des éléments de preuve pouvant aider à déterminer les « effets différentiels » des interventions. Même si des données bien comprises et bien structurées peuvent y contribuer grandement, de nombreux problèmes se posent lorsqu'on cherche à obtenir des estimations fiables des effets différentiels en contexte non expérimental. Dans beaucoup de cas, il pourra se révéler tout simplement impossible de produire des résultats valides à partir des données disponibles seulement. C'est pourquoi les évaluations planifiées doivent comprendre des procédures permettant d'analyser les points forts et les points faibles, de même qu'un processus clair pour l'intégration des résultats et leur diffusion auprès d'auditoires plus vastes.

Les mises en garde qui précèdent sont l'un des éléments qui justifient le cadre général présenté dans ce document, et elles y sont rappelées fréquemment. Elles sont nécessaires à une interprétation adéquate des suggestions du groupe d'experts. Pour interpréter et présenter adéquatement les éléments de preuve qui seront recueillis et qui pourront être diffusés à l'avenir, il faudra reconnaître les points forts et les points faibles possibles de diverses approches envisagées pour élaborer des évaluations sommatives.

---

1 Les PEMS sont des « prestations d'emploi et mesures de soutien ». C'est la terminologie utilisée à l'échelle nationale, mais le nom du programme peut varier dans les régions.

2 Ce rapport réunit, structure et développe une série de documents résumant les vues que le groupe d'experts sur les PEMS a exprimées lors d'une rencontre de deux jours (les 8 et 9 mars 2001 — voir Nicholson, 2001) et à l'occasion d'autres discussions générales ou plus pointues qui ont suivi.

Le document se divise en cinq sections analytiques qui examinent les éléments conceptuels généraux suivants :

A. Définition des participants au programme des PEMS
B. Méthodes d'analyse (y compris la sélection des groupes témoins)
C. Plan d'échantillonnage
D. Mesure des résultats obtenus par les participants
E. Intégration au projet des indicateurs à moyen terme (IMT)

Une dernière section (F) résume certains des enjeux examinés et qui devraient être considérés au moment de la conception d'évaluations quasi expérimentales et comprend certaines lignes directrices sur la mise en œuvre des évaluations des PEMS.

A. Définir les « participants » aux fins des évaluations des prestations d'emploi et mesures de soutien (PEMS)

Pour concevoir une analyse claire des répercussions probables des interventions qu'englobe le programme des PEMS, il faut d'abord définir précisément la nature du programme et la façon dont ses « participants » sont déterminés. Deux facteurs généraux ont guidé l'examen de ces enjeux par le groupe d'experts : 1) la définition du programme et de ses interventions devrait être celle qui reflète le mieux la façon dont les services sont dispensés; et 2) la définition de la participation devrait refléter les expériences des individus, sans préjuger des résultats. Dans ce contexte, le groupe d'experts a fait les observations et les suggestions suivantes :

1. La définition des participants en fonction de la date de début du plan d'action semble être celle qui répond le mieux aux objectifs de l'évaluation. Le groupe d'experts est d'avis que c'est le concept du plan d'action qui reflète le mieux l'approche globale de la « gestion des cas » qui a été retenue dans le cadre du programme des PEMS. Même si les interventions dont profitent les participants sont assez différentes, la plupart des participants suivent le même processus d'inscription au programme, étant délibérément aiguillés vers les services jugés les plus avantageux. Par conséquent, le plan d'action (même s'il n'est pas utilisé explicitement dans toutes les régions) semble correspondre le plus étroitement à la nature du « programme » selon l'idée que s'en font les décideurs.

Il a paru au groupe d'experts que l'utilisation des dates de début des plans d'action pour identifier les participants au programme présentait divers avantages. Premièrement, il semble probable que les dates de début soient définies de façon relativement claire pour tous les participants. La solution de rechange qui consisterait à se concentrer sur les personnes qui ont cessé de participer au programme au cours d'une période donnée semble poser plus de problèmes, en partie parce que les dates de fin des plans d'action sont souvent définies de façon arbitraire. Une deuxième raison de préférer la date de début pour définir la participation est qu'il est probable qu'une telle définition corresponde mieux à d'autres données publiées, par exemple dans le cadre des Rapports de contrôle et d'évaluation. Finalement, la date de début semble représenter, d'un point de vue conceptuel, une ligne de démarcation plus nette entre les activités d'avant le programme et les « résultats obtenus par les participants », car, à certaines fins, tout ce qui se produit après la date de début peut être considéré comme un « résultat obtenu par les participants ». Ainsi, la période qui s'écoule pendant le programme est naturellement incluse et les coûts d'option afférents à la participation au programme peuvent être intégrés à l'évaluation générale.

Le groupe d'experts reconnaît évidemment que certaines régions n'ont pas recours à la formule de la gestion des cas/du plan d'action pour dispenser les interventions dans le

cadre des PEMS, et que la définition des participants en fonction des dates de début posera des problèmes conceptuels dans le cas d'interventions multiples ou lorsqu'il faudra faire, du point de vue des politiques, la distinction entre les résultats obtenus pendant et après le programme. Même si le groupe d'experts est convaincu que ces inconvénients possibles n'enlèvent rien à l'intérêt de la démarche qu'il propose (et qui semble suffisamment souple pour tenir compte d'un vaste éventail de procédures utilisées concrètement dans le cadre du programme[3]), il n'en croit pas moins que les concepteurs des évaluations à titre individuel devraient explorer ces questions à nouveau, en fonction des processus véritablement utilisés dans le cadre du programme et des données disponibles.

2. En général, les évaluations devraient se concentrer sur les personnes qui ont participé à un volet des « prestations d'emploi ». Les quatre prestations d'emploi (« PE » — c'est-à-dire Subventions salariales ciblées, Aide au travail indépendant, Partenariats de création d'emplois et Développement des compétences) sont au cœur des PEMS. Elles sont également les interventions les plus coûteuses par participant. Le groupe d'experts est donc d'avis que les évaluations devraient se concentrer sur les participants à ces interventions. Les régions voudront probablement évaluer les répercussions de chacune de ces interventions séparément. Cela représentera des problèmes particuliers au chapitre de la conception des évaluations. Nous y reviendrons dans la section C ci- dessous, où nous montrerons que les modèles optimaux permettant de répondre à cet objectif qui consiste à évaluer des interventions à titre individuel pourront différer suivant les intérêts particuliers des décideurs régionaux. De plus, selon le groupe d'experts, il y aurait lieu d'envisager d'inclure un groupe distinct de clients participant uniquement aux mesures de soutien dans certaines évaluations. Cette possibilité est examinée ci-dessous.

3. La « participation » appelle une définition précise — peut-être selon le financement (si c'est possible). Le groupe d'experts se préoccupe à l'idée qu'une date de début d'une PE en particulier (ou d'un plan d'action en général) puisse avoir été attribuée à certains clients des PEMS sans qu'ils n'y aient vraiment pris part. En supposant que de tels cas sont relativement communs (il faudrait cependant recueillir des données à ce sujet), le groupe d'experts est d'avis qu'il devrait y avoir participation véritable, même minime, à une intervention. Une possibilité serait de définir la participation en fonction du financement observé qui a été consacré à un client ayant participé à l'intervention en question. Il faudra cependant procéder à d'autres recherches pour déterminer s'il existe des données sur le financement individuel pour chacune des interventions PE. Il semble probable que les définitions de la participation reposant sur le financement différeront d'une région à l'autre, parce que les façons dont ces données sont recueillies différeront elles aussi.

---

3 Par exemple, dans les régions qui n'ont pas retenu explicitement la formule de la gestion des cas/des plans d'action, les dates de début et de fin des interventions pourraient tenir lieu de dates de début et de fin des plans d'action. Pour déterminer précisément la façon de procéder à cette simulation, il faudra connaître à fond les politiques et les pratiques de mise en œuvre au niveau local. Il est possible que les définitions doivent varier d'une région à l'autre pour que les plans simulés reflètent fidèlement la façon dont le programme est véritablement dispensé. C'est une question que les comités conjoints d'évaluation et Développement des ressources humaines Canada (DRHC) devront examiner soigneusement lorsqu'ils concevront la version finale de leurs évaluations provinciales/territoriales et qu'ils feront connaître les résultats à l'échelle nationale.

4. L'achèvement du programme n'est pas un bon critère de sélection des membres de l'échantillon. Bien qu'il puisse paraître juste de faire porter l'évaluation des PEMS sur les seuls participants qui ont achevé le programme, le groupe d'experts pense que cette approche ne serait pas de mise. L'achèvement d'une PE est en soi un résultat présentant de l'intérêt. Il ne faudrait pas en faire une condition pour la sélection de l'échantillon.

5. Dans certaines régions, il conviendrait d'envisager une cellule distincte pour les clients des services d'aide à l'emploi (SAE) seulement. Le programme des PEMS consacre environ le tiers de son budget aux mesures de soutien, et celles-ci devraient être évaluées séparément. Des recherches antérieures ont montré que des interventions relativement modestes sur le plan de l'emploi peuvent afficher le meilleur rapport coût-efficacité (voir la section C). Le groupe d'experts croit donc que si l'on exclut de l'analyse tous les individus participant uniquement à une intervention MS, on risque de passer à côté d'une bonne partie de ce qui fait l'intérêt du programme global des PEMS. C'est particulièrement vrai dans les régions qui mettent l'accent sur les mesures de soutien. En outre, il semble possible qu'une cellule composée uniquement de participants aux SAE puisse contribuer à l'estimation de l'effet des interventions PE en soi, parce qu'un échantillon exclusivement SAE peut, dans certaines circonstances, se révéler un bon groupe témoin pour certaines interventions PE. De nombreuses autres évaluations ont adopté une approche à plusieurs niveaux pour définir les traitements, selon laquelle les traitements plus complexes représentent des « ajouts » aux traitements plus simples. Il se peut en fait que, dans certains endroits, les PEMS soient exploitées selon cette approche à plusieurs niveaux. Les rôles éventuels que pourrait jouer une cellule composée uniquement de participants aux SAE sont examinés dans les sections B et C; le groupe d'experts pense que la plupart des évaluations devraient envisager cette possibilité.

6. Les apprentis devraient faire l'objet d'une étude distincte. Même si la prestation de services aux apprentis est une composante importante du programme des PEMS, le groupe d'experts est d'avis que les questions méthodologiques qu'il faut régler avant de pouvoir étudier ce groupe adéquatement nécessiteraient un plan de recherche distinct. Parmi les grandes questions que soulèverait nécessairement une telle étude, on retrouverait probablement celles-ci : 1) Comment définir les périodes de participation au programme des apprentis? 2) Comment établir un groupe témoin pour les apprentis — est-il possible de trouver des individus présentant des « liens avec le marché du travail » similaires? 3) Comment les résultats obtenus par les apprentis devraient-ils être définis? Comme les réponses qui pourraient être apportées à ces questions ne correspondent pas exactement aux thèmes examinés dans les ouvrages plus généraux sur l'emploi et la formation, le groupe d'experts pense que le simple fait d'ajouter un « traitement » des apprentis dans le plan général d'évaluation sommative apporterait peu d'informations utiles. Cela détournerait en outre certaines ressources précieuses de la réalisation d'autres objectifs d'évaluation.

B. Méthodes d'analyse à utiliser dans les évaluations

Le groupe d'experts reconnaît que l'évaluation des effets différentiels des PEMS constitue un objectif opportun[4], mais que si on cherche à évaluer ces effets dans un contexte non expérimental (c'est-à-dire en faisant appel à des modèles quasi expérimentaux), il faudra relever des défis importants. La mesure de tels « effets différentiels » sous forme d'estimations définitives des répercussions d'une intervention sur les participants par rapport aux groupes témoins soulève d'importantes questions méthodologiques. Dans beaucoup de cas, compte tenu des données disponibles, il peut tout simplement être impossible de trouver des éléments de preuve définitifs au sujet des « vraies » répercussions. Les résultats des évaluations devraient alors, dans certaines conditions clairement définies, être présentés comme des « indicateurs de répercussions possibles ». L'extension des résultats à des situations qui ne correspondent pas à ces conditions devrait être soumise à un examen attentif et critique.

Les enjeux peuvent être illustrés comme suit : un objectif des évaluations sommatives est de produire des estimations « cohérentes » et relativement « précises » des résultats chez les participants et les membres du groupe témoin. En d'autres termes, la méthode devrait donner des estimations qui, si les échantillons étaient très vastes, se rapprocheraient des valeurs vraies (pour l'ensemble de la population). Les estimations devraient aussi afficher une faible variabilité d'échantillonnage, de sorte qu'il soit peu probable qu'elles s'écartent beaucoup des valeurs vraies. En fait, les gestionnaires s'attendent souvent à ce que les évaluations puissent produire des estimations cohérentes et précises de « l'effet différentiel » (c'est-à-dire l'effet du programme sur les résultats par rapport à ce qui se serait produit en l'absence de l'intervention). Le groupe d'experts reconnaît qu'il sera difficile d'atteindre ces buts an moyen d'un modèle quasi expérimental. En effet, un tel modèle comporte inévitablement le problème fondamentalement insoluble de la comparabilité des participants et des témoins, en particulier lorsque les différences entre les groupes sont liées à des caractéristiques qui ne sont pas directement observées lors de l'évaluation. Le groupe pense qu'il n'est pas possible de préciser *a priori* la « meilleure » formule qui se révélera optimale pour produire des estimations adéquates dans toutes les circonstances. Il recommande donc que les évaluateurs adoptent une démarche vaste et variée, explorant diverses méthodes, tant pendant la phase de planification que pendant la phase d'analyse de leurs travaux. De plus, le groupe d'experts est fermement convaincu que les approches retenues devraient être soigneusement étayées et faire l'objet de comparaisons critiques. Il croit également qu'il faudrait encourager les évaluateurs à déterminer la meilleure démarche à adopter.

---

[4] D'autres objectifs pourraient se révéler importants pour certains décideurs, et notamment : 1) obtenir de bonnes données sur les caractéristiques des participants; 2) étudier le processus de mise en œuvre des interventions des PEMS; et 3) mesurer les expériences « brutes » des participants après le programme. Même si la plus grande partie de ce rapport (et des délibérations du groupe d'experts) a porté sur les moyens d'obtenir des estimations des effets différentiels, ces autres objectifs peuvent se révéler assez importants dans certaines régions et ils joueront donc un rôle de premier plan dans la conception des évaluations pour ces régions.

1. Stratégies envisageables pour la mesure des résultats

Selon le groupe d'experts, diverses approches semblent envisageables pour mesurer les résultats obtenus par les participants et les non-participants aux interventions des PEMS dans le cadre des évaluations sommatives. Pour que rien ne soit négligé, le groupe a dressé une liste assez complète de possibilités, qui sont énumérées ici, chacune étant commentée.

a. **Les méthodes de répartition aléatoire (« méthodes expérimentales ») :** La répartition aléatoire demeure la « règle d'or » dans les évaluations relatives au marché du travail. Elle garantit une estimation cohérente et efficace de l'effet moyen d'une intervention sur le participant[5] et est devenue la norme de comparaison pour toutes les autres approches (voir, par exemple, Lalonde, 1986, ainsi que Smith et Todd, 2001). Compte tenu de ces avantages, le groupe d'experts est fermement convaincu qu'il conviendrait d'envisager la possibilité de recourir à une forme ou une autre de répartition aléatoire dans les évaluations sommatives. Le groupe reconnaît évidemment que le principal inconvénient de la répartition aléatoire est qu'elle n'est pas compatible avec l'objectif fondamental de l'accès universel du programme des PEMS. Dans la mesure où les personnes faisant partie d'un groupe témoin ne pourraient pas participer au programme (du moins pendant un certain temps), cela placerait les besoins de l'évaluation en contradiction avec les principes sur lesquels repose le programme. La façon habituelle de régler de tels conflits est de concevoir des interventions améliorées par rapport aux interventions normalement offertes, selon le principe voulant qu'il soit moralement moins répréhensible de refuser à un individu de participer à l'intervention améliorée. Si le financement prévu pour certaines interventions est limité, ces interventions elles-mêmes pourraient être considérées comme des « améliorations », de sorte qu'en pareil cas, la répartition aléatoire serait également justifiable. D'autres contraintes (en ce qui concerne la capacité du programme, par exemple) pourraient également servir de critères à la répartition aléatoire dans les évaluations, de manière à ne pas entraver l'admissibilité universelle aux programmes.

   Toutefois, compte tenu des constats tirés de la mise en œuvre du programme des PEMS sur le terrain, le groupe d'experts conclut de façon générale que de telles options ne sont pas fréquentes dans le cadre du programme tel qu'il est actuellement conçu. Le groupe est néanmoins d'avis que les évaluateurs devraient toujours envisager en premier lieu la possibilité de mener une évaluation reposant sur une répartition au hasard, avant de se tourner vers des solutions de second choix. Les évaluateurs devraient également indiquer les cas où les évaluations reposant sur une répartition aléatoire pourraient se révéler fort utiles lorsqu'il s'agit de clarifier des constats ambigus devant servir à la détermination des répercussions. On pourrait ainsi cerner les cas dans lesquels la répartition aléatoire pourrait être utilisée fort efficacement pour compléter d'autres recherches visant le programme des PEMS.

---

5 Il conviendrait de modifier cet énoncé si le traitement expérimental visait également les membres du groupe témoin (par exemple, si les participants étaient placés au premier rang dans les files d'attente). On trouvera un examen de certaines questions plus complexes concernant les conclusions qu'on peut légitimement tirer des modèles de répartition aléatoire dans Heckman, Lalonde et Smith, 1999.

Il faudrait aussi considérer la répartition au hasard à des fins autres que la création de concepts de traitement-contrôle additionnnels. Une gamme d'améliorations d'importance croissante aux programmes pourraient être assignées au hasard et s'ajouter à une option de service universelle. En outre, la répartition au hasard pourrait permettre de créer des « outils » utiles pour identifier la participation au programme dans le cadre d'évaluations qui s'avéreraient autrement non expérimentales. Pour réaliser ceci, on pourrait imposer des exigences de counselling additionnelles à un groupe de prestataires de l'AE sélectionnés au hasard par exemple, ce qui pourrait entraîner une variation fortuite au niveau de la formation. On pourrait réaliser un autre objectif similaire en misant sur des primes monétaires allouées au hasard. L'essentiel, c'est qu'une répartition au hasard innovatrice pourrait s'avérer utile de maintes façons en vue d'améliorer la qualité des évaluations de PEMS. Compte tenu de la validité additionnelle qu'on n'obtient souvent qu'en utilisant la répartition au hasard, le groupe d'experts appuierait avec enthousiasme des efforts si novateurs.

b. **Les méthodes non expérimentales :** De nombreuses méthodes non expérimentales ont été utilisées pour l'évaluation des programmes du marché du travail lorsqu'une répartition aléatoire est impossible. La plupart ont recours à des groupes témoins[6], dans l'espoir que les expériences des témoins représenteront fidèlement la situation dans laquelle se seraient retrouvés les participants s'ils n'avaient pas participé au programme. Nous présenterons d'abord une liste relativement simple des possibilités. Nous verrons ensuite de façon plus détaillée, dans la sous-section suivante, de quelle façon les méthodes reposant sur des groupes témoins interagissent avec les méthodes de mesure lorsqu'il s'agit de déterminer s'il est possible de produire des estimations cohérentes des résultats obtenus par les participants. Par ailleurs, comme nous le démontrons, l'efficacité des méthodes pourra varier, selon que ce sont des facteurs observables ou non observables qui expliquent les différences entre les participants et les témoins; nous utilisons donc cette distinction pour illustrer les différentes approches.

   i. **Méthodes qui tiennent compte des différences observables.** De telles méthodes sont faciles à appliquer, mais puisqu'elles ne tiennent pas compte des différences inobservables dans les déterminants de la participation au programme, la cohérence des estimations qu'elles produisent[7] demeure douteuse. Cependant, selon certains auteurs, les stratégies d'appariement peuvent se révéler très efficaces pour constituer des groupes témoins valables (Rosenbaum et Rubin, 1983, Dehejia et Wahba, 1998, 1999), de sorte que le groupe d'experts préconise qu'elles soient envisagées parmi les méthodes de mesure des résultats dans le cadre des évaluations des PEMS :

---

6 Le recours à la méthodologie avant-après n'exige pas de groupe de comparaison explicite. Les extrapolations tirées des comportements passés jouent ce rôle.

7 Il peut y avoir incohérence, par exemple, si une variable non mesurée (notamment la « motivation ») influence à la fois la participation au programme et les résultats sur le marché du travail. Si les personnes plus motivées sont plus susceptibles de participer et qu'elles obtiennent également des résultats plus favorables sur le marché du travail lorsqu'il n'y a pas de participation, la participation au programme semblera « expliquer » ces résultats favorables.

- **Comparaison des moyennes rajustées par les moindres carrés ordinaires (MCO).** Ces estimations sont généralement utilisées pour faire une analyse préliminaire des données. La méthode donne des estimations relativement valables qui tiennent compte (de façon linéaire) des différences observables entre les participants et les membres du groupe témoin. Ces résultats peuvent contribuer à clarifier la nature des données, mais ne sauraient être considérés comme des estimations définitives des répercussions, surtout parce que la méthode n'offre aucune garantie contre l'influence que peuvent avoir des différences non mesurées, comme celles qui relèvent de facteurs inobservables.

- **MCO avec variables décalées des résultats.** Ces estimations présentent des problèmes semblables à ceux des estimations simples par les MCO. Dans certains cas, on peut améliorer la situation en tenant compte des résultats décalés, ce qui permet de prendre en considération une partie des différences inobservées entre participants et membres du groupe témoin[8]. Cependant, on risque ainsi d'introduire des biais dont on ne connaît pas le sens et les estimations qui en résultent peuvent être très sensibles à la façon dont les variables décalées ont été spécifiées.

- **Méthodes reposant sur l'appariement.** Il existe diverses stratégies d'appariement qui pourraient être mises en œuvre dans les évaluations sommatives des PEMS[9]. Comme nous le verrons dans la section qui suit, elles pourraient reposer uniquement sur des données administratives, ou encore sur une combinaison de données administratives et de données d'enquête. Même si l'appariement ne tient pas directement compte des différences non mesurées entre participants et membres du groupe témoin, il peut le faire de façon approximative si les facteurs inobservables sont corrélés aux variables utilisées pour l'appariement[10]. Comme nous le verrons dans la section C, l'adoption de procédures d'appariement aurait également des conséquences pour la répartition de l'échantillon — notamment parce qu'elle aurait pour effet d'accroître la taille des groupes témoins afin de tenir compte de la possibilité que certains des membres de l'échantillon initial ne soient bien appariés à aucun participant. L'adoption de stratégies d'appariement poserait également des problèmes de logistique au moment de la collecte des données d'enquête aux fins de l'évaluation, et on y reviendra dans la section C.

---

8 Par exemple, comme la « motivation » influence les résultats antérieurs aussi bien que les résultats futurs sur le marché du travail, on peut, en tenant compte des résultats antérieurs, obtenir une mesure (imparfaite) de la motivation.

9 En général, les méthodes d'appariement peuvent être jugées supérieures aux procédures MCO lorsqu'il s'agit de contrôler les variables observables parce que cette méthode n'entraîne pas d'hypothèse linéaire implicite aux MCO.

10 Techniquement, la qualité des données utilisées au niveau de l'appariement devrait faire en sorte que la participation au programme soit indépendante des résultats non traités, conditionels aux variables utilisées pour procéder à l'appariement.

On retrouve deux formules générales d'appariement dans la littérature :

- **L'appariement aux caractéristiques:** Ces procédures utilisent soit un appariement exact entre des cellules fondé sur les caractéristiques du participant ou une sorte d'algorithme de distance qui apparie les membres du groupe témoin[11] à des participants donnés, selon des caractéristiques observables. Les estimations peuvent différer beaucoup selon les caractéristiques utilisées au niveau du concordeur. Dans certains cas, les chercheurs ont tenté de simuler les incertitudes que suppose l'appariement exact en ayant recours à plusieurs applications différentes des algorithmes d'appariement.

- **L'appariement probabiliste :** Ce procédé (dont Rosenbaum et Rubin, 1983, ont été les pionniers) effectue l'appariement à l'aide de la « propension » prédite à participer à un programme. Il faut d'abord faire une estimation de premier degré des probabilités de participation en fonction de caractéristiques observables. La procédure peut donner de bons résultats si deux hypothèses additionnelles[12] se concrétisent : 1) la participation peut être « très précisément » prédite ; et 2) la distribution des probabilités prédites est « suffisamment semblable » entre le groupe des participants et le groupe témoin. Dans la pratique, l'appariement probabiliste peut aussi poser des difficultés de mise en œuvre, parce que les déterminants de la participation ne sont pas bien connus et que des spécifications différentes de l'équation de la participation peuvent donner des appariements assez différents. (Les chercheurs devraient énumérer, expliquer et étayer soigneusement les critères d'utilisation des expressions « très précisément » et « suffisamment semblable »).

ii. Méthodes qui tiennent compte des différences inobservables

Ces méthodes reposent également sur des hypothèses au sujet de la nature des différences inobservables entre les participants et les non-participants. S'appuyant sur ces postulats, elles tentent d'expliquer et d'éliminer ces différences *a priori* par la soustraction, la modélisation complexe ou une sélection minutieuse des groupes témoins à partir de caractéristiques détaillées. La validité des estimations dépend de la qualité des données disponibles, du bien-fondé des postulats concernant les facteurs inobservables ainsi que de la robustesse des techniques. Il

---

11 Selon les deux formules d'appariement, les cas de comparaison sont choisis « avec remplacement » — c'est-à-dire qu'un cas de comparaison peut parfois être le plus étroitement apparié à deux participants ou plus. On pense généralement que la réduction d'efficacité que représente cette double utilisation est plus que compensée par un meilleur appariement. Notons cependant que cette méthode ne serait pas adéquate si le bassin des non-participants était si limité par rapport à celui des participants que de nombreux cas seraient tirés à répétition (par exemple, la taille du bassin des candidats à l'appariement ne devrait pas être de plus de deux ou trois fois inférieure à celle du groupe de participants). Généralement, les procédures d'appariement ne fonctionneront pas bien si l'ensemble des caractéristiques d'appui font état d'une grosse différence entre la comparaison et les groupes participants.

12 C'est à dire en plus de l'hypothèse générale nécessaire à toutes les procédures d'appariement à l'effet que la participation au programme doit être indépendante des résultats non traités, compte tenu des variable observables.

est donc important que les évaluateurs, au moment d'évaluer les répercussions des PEMS, prennent en compte les dérogations aux postulats sous-jacents et leurs effets possibles sur les estimations correspondantes.

Les méthodes qui tentent de tenir compte des différences inobservables incluent :

- **Les méthodes des écarts dans les différences.** Ces méthodes reposent sur l'hypothèse cruciale voulant que les différences inobservables entre les individus soient constantes dans le temps pour chaque individu — de sorte qu'elles disparaissent dans les données qui concernent chacun au moment où sont calculées les différences attribuables aux facteurs observables. Avec un nombre suffisant d'observations antérieures ou postérieures au programme, il est possible, en principe, de tester cette hypothèse en ayant recours à des périodes de temps autres pour mesurer les différences antérieures et postérieures. Cependant, si les différences inobservables changent au fil du temps ou sont influencées par la participation au programme, la méthode des écarts dans les différences ne donnera pas d'estimations cohérentes.

- **Méthodes Heckman/VI.** Ces méthodes reposent sur l'existence d'une « variable instrumentale » (VI) devant satisfaire deux critères : 1) être indépendante du résultat mesuré; et 2) avoir une valeur prédictive significative pour expliquer les décisions relatives à la participation au programme. Les « bons » instruments sont plutôt rares, et les évaluateurs qui adoptent ces techniques devraient se montrer prudents. Dans certains cas, on peut générer des instruments lors d'une évaluation[13]. Le groupe d'experts est d'avis que, lorsqu'ils présentent les résultats obtenus selon ces méthodes, les évaluateurs devraient indiquer clairement l'instrument qui a été utilisé et fournir des tests relatifs aux spécifications qui permettront de déterminer si la variable répond aux critères nécessaires. Les évaluateurs devraient également envisager la possibilité de générer des variables instrumentales, lorsque c'est possible.

2. Groupes témoins, méthodes d'estimation et cohérence

Les techniques décrites dans la sous-section précédente ont souvent été utilisées de diverses façons dans les évaluations. Dans certains cas, une procédure est adoptée *a priori* et aucune autre n'est utilisée. Dans d'autres cas, les chercheurs adoptent une démarche plus éclectique, dans le cadre de laquelle ils recourent à une méthode d'appariement, font ensuite une analyse de régression puis, pour faire bonne mesure, appliquent certaines techniques VI. En général, le groupe d'experts estime que ni l'une ni l'autre de ces démarches n'est idéale. La sélection d'une seule démarche analytique dans le cadre d'une évaluation pourra souvent se révéler inadéquate une fois les données examinées en détail. D'autre part, une démarche hétéroclite peut donner lieu à l'adoption de techniques

13 Par exemple, l'information découlant des évaluations faites par le personnel au sujet des candidats lorsqu'il s'agit de déterminer s'ils sont aptes ou non à participer à un programme peut servir de variable instrumentale de ce genre.

incompatibles les unes avec les autres. Ce qu'il faut plutôt, c'est une justification détaillée de toutes les approches, quelles qu'elles soient, qui ont été retenues pour une évaluation en particulier, parallèlement à une évaluation *a posteriori* de la validité vraisemblable des approches retenues.

Afin de clarifier les questions ayant trait à la sélection des approches analytiques aux fins des évaluations, le tableau 1 examine de façon assez détaillée l'interaction entre le choix du groupe témoin et la méthode d'estimation. On y envisage trois possibilités quant au choix du groupe témoin, selon la façon dont les membres seront appariés[14] aux participants :

1. appariement selon une série limitée de variables administratives (par exemple, les données de l'assurance-emploi (AE) seulement), dites V1. Pratiquement toutes les évaluations, peu importe leur modèle, auront recours à cette méthode d'appariement — ne serait-ce que pour comparer les dates auxquelles commencent les périodes de versement des prestations aux prestataires;

2. appariement selon des variables V1 et d'autres variables administratives (V2 — qui, en principe, comprendraient la rémunération antérieure et l'impôt sur le revenu selon l'ADRC, car elle peut se révéler l'information administrative la plus instructive). Cela supposerait un appariement assez exhaustif à partir des données administratives avant que les enquêtes soient entreprises;

3. appariement selon des variables V1, V2 et d'autres variables que l'on peut recueillir seulement dans le cadre d'enquêtes, V3 (comme des données sur la rémunération familiale récente ou le processus à l'issue duquel les répondants se sont inscrits/ne se sont pas inscrits au programme des PEMS).

Chacune de ces méthodes de sélection des groupes témoins est corrélée aux quatre approches qui pourraient être envisagées pour générer les estimations des différences dans les résultats obtenus par les participants et par les non-participants :

1. écarts dans les moyennes;
2. écarts dans les moyennes avec rajustement par les MCO;
3. écarts dans les différences (avec rajustement par les MCO);
4. rajustement par variables instrumentales (VI) (y compris les « méthodes de Heckman »).

14 Il ne s'agira pas nécessairement d'appariements par paires. On pourrait aussi inclure des méthodes où plus d'un non participant pourrait servir à bâtir une « comparaison » pour chaque participant (voir Heckman, Ichimura et Todd, 1997).

TABLEAU 1
Cohérence des estimateurs dans diverses circonstances

Appariement du groupe témoin selon	Cohérence[15] si les participants et les membres du groupe témoin diffèrent selon			
	V1 seulement	V1 ou V2	V1, V2 ou V3	V1, V2, V3 ou U[16]
Écarts simples dans les moyennes				
V1	Oui	Non	Non	Non
V1, V2	Oui	Oui	Non	Non
V1-V3	Oui	Oui	Oui	Non
Écarts dans les moyennes avec rajustement par régression* **				
V1	Oui	Oui*	Oui*	Non
V1, V2	Oui	Oui	Oui*	Non
V1-V3	Oui	Oui	Oui	Non
Estimations des écarts dans les différences				
V1	Oui	Oui*	Oui*	Oui — si U ne varie pas dans le temps. Non autrement.
V1, V2	Oui	Oui	Oui*	Oui — si U ne varie pas dans le temps. Non autrement.
V1-V3	Oui	Oui	Oui	Oui — si U ne varie pas dans le temps. Non autrement.

* Le rajustement par les MCO produira ici des estimations cohérentes, mais moins efficaces que celles qu'on obtiendrait si l'appariement avait reposé sur toutes les variables mesurables. Il importe de noter que l'appariement reposant sur toutes les variables disponibles produirait généralement des estimations cohérentes, alors que le rajustement par les MCO pourrait ne pas accomplir ceci en raison de l'hypothèse de linéarité inhérente aux MCO.

** L'emploi d'algorithmes exhaustifs d'appariement peut influencer l'efficacité des méthodes d'estimation VI.

*** On tient pour acquis que les régressions de MCO comprendraient les trois types de variables mesurables, alors que l'appariement ne serait fondé que sur les ensembles de variables identifiées dans le tableau.

15 On fait référence à la cohérence statistique comme on l'a décrite plus tôt (si les échantillons étaient très vastes, les estimations se rapprocheraient des valeurs vraies pour l'ensemble de la population).

16 U désigne des variables inobservables — il s'agit de variables qui peuvent influencer les résultats tout comme la participation au programme, mais qui ne sont pas observées, ni dans les données administratives, ni dans les données d'enquête (par exemple la motivation).

TABLEAU 1 (suite)
Cohérence des estimateurs dans diverses circonstances

Appariement du groupe témoin selon	Cohérence si les participants et les membres du groupe témoin diffèrent selon			
	V1 seulement	V1 ou V2	V1, V2 ou V3	V1, V2, V3 ou U
Estimateurs VI				
V1	Oui**	Oui**	Oui**	Oui si les VI sont bonnes et que la méthode n'est pas compromise par l'appariement[17]
V1, V2	Oui**	Oui**	Oui**	Oui si les VI sont bonnes et que la méthode n'est pas compromise par l'appariement
V1-V3	Oui**	Oui**	Oui**	Oui si les VI sont bonnes et que la méthode n'est pas compromise par l'appariement

* Le rajustement par les MCO produira ici des estimations cohérentes, mais moins efficaces que celles qu'on obtiendrait si l'appariement avait reposé sur toutes les variables mesurables. Il importe de noter que l'appariement reposant sur toutes les variables disponibles produirait généralement des estimations cohérentes, alors que le rajustement par les MCO pourrait ne pas accomplir ceci en raison de l'hypothèse de linéarité inhérente aux MCO.

** L'emploi d'algorithmes exhaustifs d'appariement peut influencer l'efficacité des méthodes d'estimation VI.

*** On tient pour acquis que les régressions de MCO comprendraient les trois types de variables mesurables, alors que l'appariement ne serait fondé que sur les ensembles de variables identifiées dans le tableau.

Pour expliquer ce tableau, nous donnerons deux exemples. Premièrement, supposons que les seules différences entre les participants et les membres du groupe témoin se retrouvent au chapitre d'éléments facilement mesurables si on dispose de données administratives complètes (c'est-à-dire qu'ils diffèrent uniquement selon des mesures V1 et V2). Dans ce cas, c'est la troisième colonne — la colonne « V1 ou V2 » — qui s'applique, et le tableau 1 montre que pratiquement toutes les procédures d'estimation donneraient d'assez bons résultats, peu importe comment les échantillons sont appariés[18]. Dans ce cas, la meilleure solution première consisterait à apparier à la fois V1 et V2, quoique la méthode des moyennes rajustées par régression et des procédures d'appariement relativement modestes

17 Ainsi, Heckman, Lalonde et Smith (1999) affirment que des estimateurs économétriques qui sont valides pour des échantillons aléatoires peuvent ne pas l'être lorsqu'ils sont appliqués à des échantillons établis par appariement (1939).

18 On tient pour acquis que les estimateurs V1 utilisent un variable instrumentale solide. Le recours à un instrument moins efficace pourrait produire des estimations incohérentes, même si la sélection à l'intérieur du programme n'est fondée que sur V1 et V2.

(VI) pourrait aussi fonctionner comme il faut. D'autre part, dans le cas plus réaliste où il existe entre les participants et les membres du groupe témoin des différences selon des dimensions non mesurées (« U »), la qualité des estimations fluctuerait. Ainsi, l'une des formules proposées à l'occasion conjugue l'appariement à partir de données administratives et des méthodes VI pour tenir compte des variables inobservables; c'est ce choix que l'on retrouve dans le coin inférieur droit du tableau 1. Le tableau nous permet de faire deux conclusions importantes au sujet de ce choix analytique :

a. les propriétés d'une méthode reposant sur un appariement partiel et des procédures VI (Heckman) ne sont pas bien comprises. Le problème fondamental vient de ce que la cohérence des méthodes VI repose sur l'hypothèse voulant que le groupe témoin représente un échantillon aléatoire provenant d'une population plus vaste de participants éventuels. Un appariement partiel avant que ne soit estimée l'équation de la participation de premier degré utilisée dans la plupart des procédures VI serait inconciliable avec cette hypothèse. Pour déterminer quelle serait l'efficacité des procédures VI avec l'appariement partiel qu'on peut faire dans le contexte des PEMS, il faudrait faire d'autres recherches. En l'absence de résultats concluants à l'effet que cette approche produise des estimateurs cohérents avec une approche particulière, le groupe d'experts serait généralement opposé à ce genre d'approches mixtes;

b. d'autres procédures — particulièrement celles qui reposent sur les écarts dans les différences — peuvent donner d'aussi bons résultats pour régler les problèmes attribuables à des différences non mesurées entre les participants et les membres du groupe témoin lorsque ces différences sont constantes dans le temps. Les procédures VI ne représentent donc pas nécessairement la meilleure stratégie dans tous les cas, même en supposant que des variables inobservables suscitent des différences importantes entre participants et membres du groupe témoin.

Ces énoncés ne s'appliquent évidemment qu'à un seul choix analytique possible. Tous les autres choix qui pourraient être proposés devraient faire l'objet d'une analyse semblable. Il est probable qu'un tel examen approfondi de diverses approches arrive à la conclusion que « tout dépend de la nature des variables inobservables ». C'est pourquoi le groupe d'experts pense que, dans la mesure du possible, il conviendrait d'examiner en détail différentes stratégies analytiques pendant la phase de conception de chaque évaluation et que *plusieurs* formules parmi les plus prometteuses devraient être retenues pour l'analyse à proprement parler.

3. Précision des estimations

Non seulement est-il souhaitable que les estimations des répercussions du programme établies dans le cadre des évaluations soient cohérentes, mais encore, pour être utiles, doivent-elles être relativement précises. En d'autres termes, l'éventail des répercussions vraies qui correspond aux estimations calculées dans le cadre d'une évaluation doit être relativement étroit. Si ce n'est pas le cas, l'éventail peut être tellement vaste qu'il sera impossible de tirer des conclusions au sujet de l'efficacité du programme — en fait, il se peut même que les données permettent de conclure de façon plausible que les

répercussions ont été nulles (voire même négatives). Il convient d'apporter quatre arguments généraux au sujet de la précision dans les évaluations. Premièrement, l'objectif de la cohérence doit, en toute logique, passer avant celui de la précision. Si une évaluation ne permet pas d'établir des estimations cohérentes, à quoi bon[19] savoir que les estimations (inexactes) ont été mesurées précisément. Deuxièmement, la précision dépend souvent de la volonté de consacrer des ressources à une évaluation. En un sens, on peut toujours « acheter » plus de précision en accroissant simplement la taille des échantillons, de sorte qu'il y aura toujours un compromis entre la précision et la volonté de payer pour obtenir l'information qu'une évaluation est censée produire. Un troisième élément concerne la relation entre la précision et le choix des estimateurs. Certains des estimateurs figurant au tableau 1 nécessiteront des estimations reposant sur des sous-échantillons de données seulement[20], d'où une réduction inévitable de la précision. Un argument connexe plus général : pour un grand nombre des estimateurs énumérés au tableau 1, il est impossible de se prononcer clairement au sujet de la précision. En fait, dans certains cas, il est même impossible de calculer de façon analytique l'incertitude associée à une estimation et il faut l'établir de façon approximative à l'aide de méthodes de simulation d'auto-amorçage.

Compte tenu des complexités qui surgissent lorsqu'il s'agit de se prononcer sur la précision des estimations établies selon des méthodes non expérimentales, il est devenu courant d'utiliser des estimations de précision indicatives reposant sur des modèles expérimentaux dans la conception des évaluations. Ces indications tiennent implicitement pour acquis que les résultats obtenus par des méthodes expérimentales sont les plus précis qu'on puisse obtenir dans le cadre d'une évaluation, et les plans reposent sur ce scénario de la meilleure éventualité. Le tableau 2 donne un exemple de cette procédure. Il montre l'amplitude de la différence qui pourrait être décelée avec une efficacité de 80 p. 100, en supposant qu'un test de signification unilatéral à un seuil de 95 p. 100 soit utilisé dans les évaluations. Les écarts-types utilisés dans ces calculs proviennent de l'étude de faisabilité réalisée en Nouvelle-Écosse pour le projet indicateurs à moyen terme (IMT). Il se peut qu'ils sous-évaluent le niveau réel de précision qu'on pourrait obtenir lors des évaluations sommatives, car celles-ci pourraient compter sur de meilleures variables de contrôle et de meilleures méthodes statistiques que celles qui ont été utilisées pour l'étude de la Nouvelle-Écosse. Cependant, comme nous le verrons dans la section C, les estimations de la précision figurant dans le tableau correspondent approximativement aux constats de nombreuses évaluations expérimentales des programmes actifs du marché du travail.

---

19 En fait, dans certains cas, il pourrait être possible d'imposer des limites aux biais inhérents à une technique d'estimation en particulier, et on pourrait opter pour un compromis entre des effets estimés avec précision accompagnés de biais limités et des estimations cohérentes estimées de façon imprécise. Toutefois, en général, les causes des manques de cohérence ne sont pas suffisamment bien comprises pour qu'on puisse faire de tels compromis globaux.

20 C'est particulièrement vrai des estimateurs d'appariement qui reposent sur des données d'enquête et suivant lesquels certains membres des groupes témoins qui ne sont pas appariés aux participants seraient éliminés de l'analyse.

TABLEAU 2 Différences décelables, efficacité de 80 % et seuil de 95 %				
Taille des échantillons		**Résultats obtenus par les participants (ET présumé)**		
Participants	Témoins	Rémunération selon les T4 (11 000)	Occupent un emploi (0,370)	Semaines d'AE (9,00)
2 000	1 000	1 103	0,037	0,90
2 000	2 000	901	0,030	0,74
3 000	1 500	901	0,030	0,74
4 000	2 000	780	0,026	0,64
1 000	1 000	1 274	0,043	1,04
1 000	400	1 685	0,057	1,38
500	400	1 911	0,064	1,56

En tenant pour acquis que ces chiffres représentent raisonnablement les résultats qu'on peut attendre des évaluations, plusieurs observations peuvent être faites :

1. **Il conviendrait de prévoir des nombres suffisants d'observations dans les groupes témoins au moment de la conception.** En effet, en comparant les deuxième et troisième rangées du tableau, par exemple, on constate que des échantillons équilibrés 2 000/2 000 donnent exactement le même degré de précision que des échantillons déséquilibrés 3 000/1 500. Aux fins de l'estimation des différences entre les résultats des participants et ceux des témoins, 500 personnes ont été interviewées en vain dans le second cas.[21] Les deux dernières rangées témoignent également de l'importance que revêt la taille du groupe témoin. Elles illustrent les cas où l'appariement à l'aide de données d'enquête entraîne une réduction importante de la taille effective des échantillons. Il appert que dans de tels cas, si un programme a des répercussions modestes, elles ne seront vraisemblablement pas décelées à des seuils acceptables de signification et d'efficacité.

2. **Les concepteurs des évaluations devraient examiner soigneusement s'il sera possible de déceler des différences suffisamment petites entre les témoins et les participants à des interventions particulières.** Les calculs d'efficacité montrent que seuls les effets très prononcés seront décelables, étant donné la taille prévue des échantillons aux fins de l'analyse de diverses interventions de prestations d'emploi(PE). Parallèlement aux problèmes éventuels qui peuvent se poser dans la sélection des groupes témoins pour des interventions données, ces constats nous amènent à penser que les attentes relatives à la mesure des répercussions

21 Ces plans déséquilibrés peuvent s'avérer utiles si le coût des cas de participants et le coût des cas témoins varient considérablement. Mais dans la plupart des évaluations non expérimentales, tel n'est pas le cas. Même au niveau des évaluations expérimentales, on remarque habituellement que l'observation des participants est plus coûteuse que celle des groupes témoins. Ainsi, des concepts déséquilibrés devraient inclure des groupes témoins proportionnellement plus gros.

d'interventions individuelles doivent être assez modestes, tout comme les possibilités d'estimer les répercussions pour d'autres sous-groupes ou à l'intérieur d'une province.[22]

3. **Il conviendrait d'envisager des échantillons plus vastes pour les évaluations.** Même si les chiffres du tableau indiquent que la plupart des différences entre participants et non-participants pouvant témoigner du rapport coût-efficacité de l'intervention devraient être décelables, compte tenu des tailles d'échantillon prévues (on envisage environ 3 000 entrevues), certains des chiffres, particulièrement ceux qui concernent les améliorations au chapitre de la rémunération, posent des problèmes. Les différences décelables pourraient facilement devenir déraisonnablement grandes si des préoccupations relatives à la validité donnent lieu à l'élimination de vastes portions de certains des groupes témoins. On devrait examiner dans quelle mesure des échantillons de plus grande taille permettraient d'atténuer ces problèmes.

4. **Il conviendrait de faire un suivi attentif des taux de réponse dans les évaluations.** Compte tenu des nombreux problèmes qui entourent la validité des approches analytiques dans les évaluations, il faudrait déployer tous les efforts possibles pour s'assurer que les données ne sont pas entachées de problèmes plus banals (dont certains seront examinés dans la prochaine section). Les données administratives disponibles devraient permettre de calculer l'ampleur de la non-réponse dans les évaluations et on devrait s'attendre à une analyse de cette question dans le cadre des évaluations.

---

22 Étant donné les échantillons beaucoup plus vastes inclus dans les données administratives qui doivent servir pour le projet des IMT, on peut penser que ce dernier est l'instrument tout indiqué pour l'examen des répercussions dans les sous-groupes.

C. Plan d'échantillonnage

Il faut répondre à trois grandes questions au moment de concevoir les échantillons qui seront utilisés dans les évaluations sommatives : 1) Comment le groupe des participants sera-t-il sélectionné? 2) Comment les données administratives et les données d'enquête seront-elles utilisées de concert pour la sélection du ou des groupes témoins? 3) Comment les ressources devraient-elles être réparties en fonction des échantillons?[23]

1. Sélection des participants :

Selon le groupe d'experts, le tirage d'un échantillon de participants pour les évaluations ne pose pas de graves problèmes conceptuels (en supposant que les données administratives soient de qualité suffisante). Les participants seraient sélectionnés à partir des données administratives, selon la méthode qui représente le mieux l'activité du programme pendant une période donnée. Le groupe d'experts a fait trois recommandations mineures au sujet de ce processus de sélection :

a. l'échantillon des participants devrait s'étaler sur une année complète[24], de façon à minimiser les influences saisonnières qui peuvent se manifester dans les résultats;

b. il conviendrait de stratifier l'échantillon selon le genre de prestations d'emploi (PE) et la région. On assurerait ainsi une représentation adéquate des interventions comptant relativement peu de participants. Cette formule permettrait également de tenir compte des variations qu'on peut retrouver dans le contenu du programme d'une région à l'autre;

c. l'échantillon devrait être tiré de façon à ce qu'il se soit écoulé un certain temps entre la fin du programme et le début de l'enquête. En termes concrets, au moins une année devrait s'être écoulée entre la fin de la période visée par l'échantillon et le début de la période d'enquête[25].

---

23 On considère ici que les données administratives sont abondantes et qu'il n'en coûte rien de les recueillir. La taille des échantillons pendant la phase de collecte des données administratives des évaluations est donc considérée comme étant illimitée. Cependant, pour certaines interventions peu fréquentes, ce ne sera peut-être pas le cas, de sorte que nous discutons ici brièvement de la répartition de l'échantillon parmi les interventions.

24 Si on retenait l'exercice financier, cela faciliterait également les comparaisons par rapport à d'autres données administratives — particulièrement si ce sont les dates de début qui sont utilisées pour définir la participation.

25 Si les enquêtes étaient menées sur une période d'un an, deux années se seraient alors écoulées depuis la date de début du programme. Si les enquêtes sont regroupées pour des fins d'efficacité des entrevues, le groupe d'experts recommande un délai plus long entre la fin de la période visée par l'échantillon et le début des entrevues (peut-être 18 mois ou plus).

2. Sélection du groupe témoin

La sélection d'un groupe témoin est l'un des éléments les plus importants de la conception d'une évaluation. Pour simplifier l'examen de cette question, nous tenons pour acquis que toute l'analyse sera fondée sur des estimations moyennes des répercussions des prestations d'emploi et mesures de soutien (PEMS) sur les participants avec rajustement par régression. En d'autres termes, nous écartons pour l'instant certains des problèmes relatifs aux estimations que nous avons examinés dans la section B, pour nous concentrer explicitement sur les problèmes de sélection du groupe témoin. La question de savoir si les conclusions que nous présentons ici seraient différentes si d'autres stratégies d'estimation étaient utilisées (par exemple la méthode des écarts dans les différences ou des procédures VI) devra faire l'objet d'autres recherches, comme nous le verrons dans la section F.

On dispose de trois sources possibles de données pour la sélection du groupe témoin : i) les données administratives concernant l'assurance-emploi (AE); 2) les données administratives de l'Agence des douanes et du revenu du Canada (ADRC) sur la rémunération et sur d'autres éléments relatifs à l'impôt; et 3) des données d'enquête. Il est important de comprendre les avantages et les inconvénients de ces diverses sources et comment ils peuvent influencer d'autres aspects de l'analyse.

a. **Données concernant l'AE :** Ces données sont les plus facilement accessibles pour la sélection du groupe témoin. Les témoins pourraient être choisis selon leurs antécédents en matière d'AE et/ou selon des données concernant des cessations d'emploi antérieures (les données des relevé d'emploi (RE). Cette procédure donnerait probablement des résultats assez médiocres pour ce qui est de l'appariement par rapport aux antécédents des participants en matière d'emploi. Ce serait particulièrement vrai dans le cas des anciens prestataires réadmissibles de l'AE — ceux pour lesquels il n'y a pas actuellement de demande active de prestations. Même s'il y avait moyen de tirer un échantillon témoin de personnes ayant présenté des demandes de prestations dans le passé, celles-ci auraient vraisemblablement une expérience d'emploi plus récente que les anciens prestataires réadmissibles. Partant, même si l'appariement reposant uniquement sur les données de l'AE était considéré comme une méthode adéquate pour le groupe des prestataires actifs (ce qui n'a pas été démontré), il faudrait adopter des procédures supplémentaires pour les anciens prestataires réadmissibles.

b. **Données sur la rémunération de l'ADRC :** Les données sur la rémunération provenant de l'ADRC peuvent jouer un rôle essentiel dans la conception des évaluations sommatives. On sait[26] que les tendances de la rémunération pendant la période précédant immédiatement un programme sont un prédicteur important de la participation (Ashenfelter, 1978, Heckman et Smith, 1999). Plus généralement, on

[26] C'est du moins un constat fréquent dans les évaluations des programmes actifs du marché du travail aux États-Unis. Heckman, Lalonde et Smith (1999) suggèrent que ce phénomène de « baisse Ashenfelter » pourrait bien être mondial. L'ampleur du phénomène chez les participants aux PEMS n'est pas connue, mais certains travaux préliminaires portant sur des données recueillies en Nouvelle-Écosse (Nicholson, 2000) portent à croire que cette baisse survient aussi dans le cas de ce programme.

pense qu'une méthode permettant de tenir compte adéquatement des tendances de la rémunération serait un moyen prometteur pour régler les problèmes d'évaluation posés par les variables inobservables (Ashenfelter, 1979). Cette hypothèse est confirmée par certaines estimations préliminaires des répercussions des interventions des PEMS en Nouvelle-Écosse (Nicholson, 2000), qui montrent comment les données de l'ADRC peuvent être utilisées pour sélectionner un groupe témoin général de façon à ce qu'il ressemble davantage au groupe des participants[27]. Malheureusement, on ne sait pas encore dans quelle mesure les évaluateurs des PEMS auront accès à ces données. Mais compte tenu de la méthode de sélection proposée et du calendrier de réalisation de l'enquête, il serait possible, selon les normes précédentes d'accessibilité, d'établir un profil exhaustif de la période antérieure au programme pour pratiquement tous les participants et les témoins éventuels. Peu importe qu'on choisisse une méthode générale pour sélectionner un groupe témoin ou qu'on ait recours à une méthode d'appariement par paires à titre individuel, il semble fort probable qu'on puisse obtenir une concordance assez rapprochée au chapitre de la rémunération observée.

c. **Données d'enquête :** Troisième source possible de données pour la sélection du groupe témoin : l'enquête de suivi qui serait menée environ deux ans après l'inscription au programme des PEMS. L'avantage de cette source, c'est qu'elle donne l'occasion de recueillir des données cohérentes et très détaillées auprès des participants et des témoins éventuels au sujet de leur activité sur le marché du travail avant le programme, et d'autres données ayant trait à l'inscription éventuelle à des interventions PEMS. Ces données peuvent être utilisées dans le cadre de différentes méthodes (y compris l'appariement selon les caractéristiques et la propension, et certaines procédures VI) permettant de vérifier les différences au niveau des participants/comparaisons.

   Les avantages éventuels des données d'enquête au moment de définir les méthodes analytiques qui seront utilisées dans le cadre des évaluations ne devraient cependant pas faire oublier les inconvénients de ces données, et notamment :

   - Des données d'enquête sur les activités préalables au programme ne représenteront pas une base de référence véritable. On recueillera plutôt les données à l'aide de questions qui demanderont aux répondants de se souvenir d'événements qui ont eu lieu il y a plusieurs années. Les **erreurs de mémoire** dans de telles enquêtes peuvent être très importantes — et elles se répercuteront directement sur les méthodes qui comptent beaucoup sur ce genre de données d'enquête rétrospectives.

   - L'utilisation de données d'enquête pour sélectionner les groupes témoins se traduira nécessairement par une certaine **réduction de la taille des échantillons** qui serviront en fin de compte à l'analyse — tout simplement parce qu'il pourra

27 D'autres variables provenant des données fiscales de l'ADRC pourraient servir à l'appariement des participants et des témoins, par exemple le revenu total, le revenu familial total, le nombre de personnes à charge, etc.

arriver que certains répondants ne fassent pas l'affaire comme membres du groupe témoin. L'ampleur de cette réduction dépendra beaucoup de la portée de l'appariement qui peut être fait à partir des données administratives. En l'absence de données provenant de l'ADRC, ces réductions pourraient se révéler considérables. Ainsi, une grande partie des fonds consacrés à l'enquête pourra, en fin de compte, avoir été dépensée sans servir à l'analyse.

- Finalement, il se peut que l'utilisation de données d'enquête pour la sélection des groupes témoins **compromette les données primaires sur les résultats** qui doivent être recueillies dans le cadre de l'enquête. De toute évidence, ce problème se produirait s'il faut consacrer tellement de temps à recueillir des informations exhaustives relatives à la période d'avant le programme dans le cadre de l'enquête que la collecte de données plus détaillées relatives à la période d'après le programme risque d'en souffrir. En outre, et plus subtilement, si on recueille des données exhaustives visant la période d'avant le programme et la période d'après le programme dans le cadre d'une même enquête, il se peut que les répondants adaptent leurs réponses de telle façon que des biais latents s'y manifestent. Même si cette possibilité ne repose pas sur des preuves empiriques solides, le personnel qui effectue l'enquête vérifiera souvent par en avant et par en arrière un ensemble de réponses spécifiques pour voir s'il y a cohérence. Ceci peut améliorer la qualité des données mais également biaiser la déclaration des données rétrospectives et, par conséquent, certaines estimations d'impact.

3. Approches proposées

S'agissant du problème de la spécification des groupes témoins, on peut envisager deux démarches :

a. **L'approche idéale :** Il semble clair que, compte tenu des données qu'il y aurait moyen d'obtenir aux fins des évaluations, l'approche idéale pour la sélection du groupe témoin serait d'utiliser à la fois des données de l'AE et des données de l'ADRC pour constituer un échantillon témoin qui se rapproche étroitement de l'échantillon des participants, selon des critères observables dans ces deux sources de données. L'appariement par paires pourrait être possible au moyen de ces vastes fichiers de données administratives, mais ne sera peut-être pas la meilleure méthode dans une perspective globale, selon les stratégies analytiques précises qui doivent être mises en œuvre. Par exemple, une stratégie axée davantage sur la sélection d'un échantillon à des fins d'appariement pourrait se révéler préférable dans les applications où on se propose d'utiliser des procédures d'estimation VI, parce que la sélection peut retenir plus de la variance sous-jacente au niveau des variables qui déterminent la participation. L'appariement par paires pose également certains problèmes logistiques de mise en œuvre à l'étape de l'enquête. On pourrait faire des enquêtes auprès d'un **échantillon aléatoire de paires**, mais la non-réponse de la part de l'un ou l'autre des membres d'une paire aggraverait considérablement le biais de non-réponse. Bien sûr, aucun mode d'appariement fondé sur des variables observables ne peut garantir que des différences non observables n'entraîneront pas,

en bout de ligne, des résultats incohérents. Il reste qu'une approche ayant recours à des données de l'AE et de l'ADRC pour la sélection du groupe témoin semblerait la plus prometteuse dans bien des cas. Par contre, il serait nettement avantageux de pouvoir recourir à certaines recherches sur la meilleure façon d'utiliser l'appariement de première étape au niveau d'une évaluation qui reposerait, au bout du compte, sur des données d'enquête.

b. **L'approche de rechange :** Si on n'a pas accès aux données de l'ADRC à des fins de sélection de l'échantillon dans le cadre d'une évaluation, il faudrait se rabattre sur diverses procédures qui sont nettement de second choix. On commencerait par un appariement approximatif à partir des données accessibles de l'AE. Les sources de données et les variables à utiliser pourraient varier selon les évaluations, d'après la nature des marchés du travail locaux et les participants aux interventions PE. Après cet appariement approximatif, toutes les autres procédures relatives aux groupes témoins reposeraient sur les données d'enquête. Cette approche aurait trois grandes conséquences pour la conception globale des évaluations :

   i. **l'enquête devrait être plus longue,** pour permettre de recueillir une information adéquate sur les activités au sein du marché du travail avant le programme;

   ii. **il faudrait un groupe témoin plus nombreux** par rapport à l'approche « idéale » (voir la sous-section suivante), pour tenir compte de la présence possible de sujets qui finiraient par se révéler non comparables;

   iii. **le rôle relatif des méthodes d'appariement devrait être réduit** dans les évaluations (ne serait-ce qu'en raison de la réduction de la taille des échantillons qui est inévitable lorsqu'on a recours à une méthode d'appariement reposant sur des données d'enquête) et le rôle des procédures VI[28] devrait être élargi. En raison de la nécessité de recourir aux procédures VI, il peut s'avérer nécessaire de consacrer des ressources additionnelles au développement de « bons » instruments, soit en procédant à la collecte de données additionnelles[29] ou en engageant un processus quelconque de répartiton au hasard.

4. Taille des échantillons

Vu les difficultés et les incertitudes liées à la conception des évaluations sommatives, il est difficile de se prononcer définitivement sur la taille souhaitable des échantillons. Le groupe d'experts est néanmoins convaincu qu'à ce sujet, on peut tirer certaines conclusions d'autres évaluations — particulièrement celles qui ont fait appel à la

---

28 On pourrait également utiliser de façon plus exhaustive les méthodes des écarts dans les différences, bien que leur application à des données provenant d'une seule enquête ouvre la possibilité de corrélations dans les erreurs de déclaration qui risquent de biaiser les résultats au fil du temps.

29 La cueillette de données sur une étendue géographique pour les fournisseurs de services, par exemple, pourrait s'avérer un instrument utile dans le cas de certains emplacements.

répartition aléatoire. Comme, en principe, les échantillons répartis de façon aléatoire ne posent pas de problèmes analytiques particuliers, on peut les considérer comme des « scénarios de référence » à partir desquels les modèles non expérimentaux peuvent être jugés. Dans des conditions idéales, où tous les postulats se vérifient, un modèle non expérimental serait l'équivalent d'une expérience avec répartition au hasard, une fois appliquées les méthodes analytiques appropriées (par exemple, l'appariement par paires ou des techniques VI). Ainsi, les expériences avec répartition aléatoire qui ont déjà été menées constituent un modèle intéressant pour les évaluations.

Le tableau 3 indique la taille des échantillons utilisés pour l'analyse[30] de quelques-unes des grandes évaluations à répartition aléatoire[31] menées aux États-Unis.

TABLEAU 3
Taille des échantillons de certaines études à répartition aléatoire

Évaluation	Taille de l'échantillon expérimental	Taille de l'échantillon témoin	Nombre de traitements
National JTPA	13 000	7 000	3
Illinois UI Bonus	4 186	3 963	1
NJ UI Bonus	7 200	2 400	3
PA UI Bonus	10 700	3 400	6
WAUI Bonus	12 000	3 000	6
WA Job Search	7 200	2 300	3
SC Claimant	4 500	1 500	3
Supported Work	3 200	3 400	1
S-D Income Maint.	2 400	1 700	7
National H.I.	2 600	1 900	3

Diverses tendances ressortent de ce tableau sommaire :

- les échantillons sont tous d'assez grande taille — les groupes témoins comportent au moins 1 500 personnes et, plus généralement, 2 000 personnes ou plus;
- les expériences qui prévoient un seul traitement optent généralement pour une répartition égale entre le groupe expérimental et le groupe témoin[32];
- les évaluations comportant de multiples traitements attribuent des portions relativement plus vastes de leurs échantillons aux catégories de traitement; toutefois, les groupes témoins sont généralement plus vastes que n'importe quelle cellule à traitement unique;

30 La taille finale de ces échantillons tient compte de la non-réponse à l'enquête ou à certaines de ses questions. Il conviendrait d'accroître la taille des échantillons initiaux pour tenir compte de cette érosion.

31 Pour obtenir un résumé de plusieurs de ces évaluations, ainsi qu'un grand nombre de références, voir Greenberg et Shroder (1997).

32 Ce qui minimise la variance de l'effet estimé de traitement pour un budget d'évaluation donné, en supposant que les cas de traitement et les cas témoins sont tout aussi coûteux.

- même si ce n'est pas apparent dans le tableau, un grand nombre des évaluations ont eu recours à une méthode de traitement « étagée », qui permettait de créer des traitements plus complexes en ajoutant des composantes à des traitements simples (ce qui a été le cas de bien des évaluations concernant l'assurance-chômage, par exemple). Dans ce cas, les traitements simples peuvent faire office de « contrôles » pour les traitements plus complexes, car la formule permet de mesurer les effets différentiels des nouveaux traitements[33]. Ainsi, le nombre réel de « contrôles » peut être sous-déclaré dans le tableau pour ces évaluations.

Comme un grand nombre de ces évaluations cherchaient à mesurer des résultats assez semblables à ceux qui doivent être mesurés dans les évaluations des PEMS, les tailles de ces échantillons sembleraient pertinentes dans le cas des PEMS. Plus précisément, on peut conclure de ces expériences que pour être efficaces, les groupes témoins doivent se composer d'au moins 2 000 personnes[34]. C'est la nature des traitements à examiner dans l'évaluation des PEMS qui justifie des groupes témoins aussi vastes en termes relatifs. Comme les cinq grandes interventions offertes en vertu du programme des PEMS dans les régions sont assez différentes les unes des autres, il ne sera pas possible de réaliser des gains d'efficience grâce aux modèles étagés caractéristiques des expériences relatives à l'AC[35]. L'hétérogénéité des caractéristiques des participants aux interventions PEMS individuelles est une autre raison de recourir à un vaste groupe témoin. Lorsqu'il s'agit d'évaluer des interventions à titre individuel, il est probable que seule une petite partie du groupe témoin puisse être utilisée pour chaque intervention, ce qui milite également en faveur d'un vaste groupe témoin.

L'évaluation d'interventions PE particulières soulève d'autres complications au moment de la répartition des échantillons. Dans la plupart des régions, ce sont les interventions DC qui sont de loin les plus courantes. Une méthode simple de sélection au hasard des interventions risquerait donc de se traduire par des échantillons de taille très réduite pour les autres interventions. C'est pourquoi il est probable que la plupart des évaluations recourent à des échantillons de participants stratifiés selon l'intervention et accordent plus de poids aux interventions plus restreintes[36]. Si une évaluation recourt à cette structure, le choix d'une structure de comparaison devient complexe. Le groupe témoin devrait-il lui aussi être stratifié de façon à ce que chaque sous-groupe corresponde aux participants à

---

33 Dans un grand nombre des évaluations, toutefois, les traitements moins élaborés se révèlent souvent les plus efficaces. C'est le cas de pratiquement toutes les évaluations expérimentales ayant trait à l'AC.

34 Cette conclusion est grosso modo conforme aux calculs indicatifs de l'efficacité de la section B (qui eux-mêmes reposent sur les variances observées dans les données de la Nouvelle-Écosse). Précisons à nouveau que les tailles suggérées ici sont celles des échantillons définitifs. En d'autres termes, il s'agirait du nombre de cas pouvant effectivement être analysés, une fois éliminés les cas inadmissibles, la non-réponse aux enquêtes, et ainsi de suite. De plus, lorsqu'il faut recueillir des données sur le revenu, il serait bon que les enquêtes sur le terrain soient menées à l'époque des déclarations d'impôt, pour que les répondants aient ces renseignements sous la main et en mémoire pendant l'entrevue.

35 Toutefois, on pourrait envisager un modèle étagé en intégrant une cellule concernant les SAE seulement à certaines évaluations. Les expériences relatives à l'assurance-chômage aux États-Unis laissent penser qu'un traitement concernant exclusivement les SAE pourrait effectivement présenter certains effets décelables.

36 Plus généralement, l'échantillon des participants pourrait être réparti entre les interventions de référence en fonction des intérêts des décideurs régionaux. Le plan d'attribution pourrait donc représenter un compromis entre la représentation proportionnelle et la stratification en cellules de taille égale.

telle ou telle strate d'intervention? Ou devrait-il être établi en fonction des caractéristiques d'un échantillon aléatoire de participants? Cette dernière formule pourrait être préférable si les procédures VI devaient jouer un rôle de premier plan dans l'analyse, puisqu'un groupe de comparaison plus large pourrait aider à prévoir l'affectation aux traitements. D'autre part, le recours à des groupes de comparaison propres à l'intervention pourrait s'avérer préférable si les procédures d'estimation sont plus simples. On devrait s'attendre à ce que les évaluateurs examinent soigneusement ces enjeux lorsqu'ils conçoivent les évaluations.

5. Questions relatives à la mise en œuvre

La mise en œuvre de toute stratégie d'appariement dans le cadre d'une évaluation qui compte sur des données d'enquête pose des problèmes particuliers. L'objectif général est d'éviter de « gaspiller » des entrevues auprès de membres du groupe témoin qui ne seront jamais utilisées ultérieurement dans l'analyse. Pour cela, il faut généralement qu'une bonne partie des entrevues auprès des participants ait été réalisée avant les entrevues auprès des témoins. De cette façon, les caractéristiques des membres du groupe témoin peuvent être rajustées en fonction de l'échantillon des participants[37]. Une stratégie de sondage encore plus complexe pourrait se révéler nécessaire si on ne dispose que de données administratives très limitées pour la sélection de l'échantillon. En pareil cas, il pourra se révéler nécessaire de prévoir certaines questions d'élimination dans les questionnaires des membres du groupe témoin, de sorte qu'on puisse mettre fin aux entrevues des cas non comparables avant d'y avoir consacré trop de temps. Cette procédure a été suggérée, par exemple, pour régler le problème de la sélection d'un groupe témoin pour les participants du groupe des anciens prestataires réadmissibles. Comme de telles procédures de sélection d'un échantillon reposant sur une enquête peuvent comporter des biais, il convient de faire preuve d'une extrême prudence si on les adopte.

37 L'adaptation de procédures de ce genre pour régler le problème de la non-réponse peut être très délicate, particulièrement si les taux de non-réponse ne sont pas les mêmes dans le groupe des participants et le groupe témoin. C'est une autre raison qui explique pourquoi les enjeux relatifs à la non-réponse méritent une place de premier plan dans la conception des évaluations.

D. Spécification et mesure des résultats obtenus par les participants

Les réflexions du groupe d'experts au sujet de la spécification et de la mesure des résultats reposent sur quatre critères : 1) il semble probable que pour la plupart des évaluations, certaines des principales mesures[38] porteront sur l'emploi après le programme[39]; 2) il faudrait donc recueillir suffisamment d'information relative à l'emploi pour que diverses mesures puissent être établies — ce qui facilitera la conception des mesures des résultats en fonction des objectifs particuliers des programmes[40]; 3) il conviendrait de recueillir des données sur diverses autres variables socioéconomiques importantes — qui serviront surtout de variables de contrôle dans l'étude des résultats en matière d'emploi; et 4) les techniques utilisées pour la collecte des données d'enquête devraient viser une comparabilité directe entre diverses évaluations régionales.

Voici sept mesures concrètes conformes à ces critères :

1. **Sauf considérations particulières, l'enquête de suivi devrait avoir lieu au moins deux ans après la date de début du plan d'action.** De la sorte, il serait plus probable que le plan d'action soit achevé bien avant la date de l'entrevue. Comme, selon les premières évaluations, les entrevues devraient avoir lieu à l'automne 2001, elles devraient viser les plans d'action dont la date de début tombe au cours de l'exercice 1999 (entre le 1er avril 1998 et le 31 mars 1999). Les évaluations dont les entrevues auront lieu plus tard pourraient se concentrer plutôt sur l'exercice 2000.

2. **Lorsque c'est possible, il conviendrait d'utiliser dans toutes les évaluations des questions semblables sur les antécédents des répondants en matière d'emploi.** Comme c'est l'emploi après le début de l'intervention et après le programme qui fera probablement l'objet de la plupart des évaluations des PEMS, il semble clair qu'une bonne partie des ressources des enquêtes devrait être consacrée à sa mesure. Des études antérieures ont montré que l'utilisation d'instruments différents pour la collecte des données peut mener à des résultats estimatifs assez différents (Heckman, Lalonde et Smith, 1999). Pour éviter cette source de variance, il conviendrait d'encourager les évaluateurs à utiliser les mêmes batteries de questions. Dans la même veine, les routines d'épuration des données et de

---

38 L'achèvement du programme, les prestations d'assurance-emploi (AE) touchées et l'inscription subséquente à d'autres interventions des prestations d'emploi et mesure de soutien (PEMS) pourraient constituer d'autres mesures.

39 En réalité, l'emploi devrait généralement être mesuré à partir de la date d'inscription au programme (et d'une date semblable pour les groupes témoins) puisque cela procure l'évaluation la plus précise des occasions d'emploi abandonnées relativement à la participation au programme.

40 En réalité, l'emploi peut être mesuré à partir de la date d'inscription au programme si elle se révèle le meilleur point de référence pour l'échantillon des participants et pour celui du groupe témoin.

construction des variables devraient être coordonnées d'une évaluation à l'autre. Les évaluateurs devraient également envisager la meilleure façon d'exploiter la documentation appartenant aux répondants (par exemple, les documents fiscaux) pour la collecte de ces données.

3. **Il conviendrait d'élaborer un certain nombre de mesures de l'emploi.** Il s'agit ici de veiller à ce que les mesures des résultats utilisées dans les évaluations correspondent bien aux objectifs prévus du programme. Même si l'on peut s'attendre à ce que toutes les évaluations élaborent les mêmes mesures de base (par exemple, le nombre de semaines d'emploi au cours de l'année écoulée, la rémunération globale pendant cette période, le nombre d'emplois occupés, etc.), il y aurait par ailleurs une certaine flexibilité pour que les régions puissent se concentrer sur les mesures qu'elles jugent les plus appropriées, compte tenu de l'éventail des prestations offertes. Par exemple, pour les clients des interventions de la composante Développement des compétences, les résultats pourraient se concentrer sur les augmentations salariales après le programme ou sur les changements dans le genre des emplois occupés. Pour les participants du volet Subventions salariales ciblées, les évaluations pourraient se concentrer sur les transitions réussies vers des emplois non subventionnés. Et il peut se révéler très difficile de concevoir des moyens de mesurer la viabilité à long terme des options de travail indépendant choisies par certains clients. Comme on l'a vu précédemment, il faudra de toute évidence faire d'autres recherches pour déterminer précisément les liens qui seront établis entre les résultats et les interventions. À un niveau plus conceptuel, il faudra montrer explicitement comment les résultats qui doivent être mesurés dans le cadre des évaluations correspondent aux objectifs sociétaux généraux du programme des PEMS (par exemple, tels qu'ils sont énoncés dans la loi habilitante).

4. **Il conviendrait d'élaborer un module de base pour la collecte de données sur d'autres variables socioéconomiques qui pourrait être utilisé dans la plupart des évaluations.** Il s'agirait ici de s'accorder sur les variables intermédiaires à mesurer et de veiller à ce qu'elles soient examinées dans toutes les évaluations. Faute d'un tel accord, il pourra être très difficile de comparer les résultats des analyses d'une région à l'autre, parce que les analyses porteraient sur différents facteurs dans différentes évaluations. La mise en commun des données à des fins d'analyse entre les régions serait également entravée si les évaluations ne comptaient pas ce module commun. De toute évidence, de telles analyses comparatives entre les régions présentent un intérêt direct pour Développement des ressources humaines canada (DRHC), parce qu'elles permettront à la fois d'accroître la précision de certaines estimations et de faire ressortir certains déterminants importants des divers degrés de réussite entre les régions[41]. Par conséquent, il faudrait envisager des moyens d'encourager tous les évaluateurs à utiliser des modules de base semblables — peut-être en chargeant un entrepreneur de les élaborer en vertu d'un contrat distinct.

---

41 On reviendra sur l'intérêt de mettre en commun les données de différentes évaluations dans la section F.

5. **Il conviendrait d'envisager d'autres entrevues de suivi, au moins dans certaines régions.** Même si la plupart des évaluations auront probablement recours à la formule des enquêtes uniques, le groupe d'experts est convaincu qu'il conviendrait d'encourager les évaluateurs à examiner ce que des suivis ultérieurs pourraient révéler (peut-être 24 mois après l'enquête initiale). Il est probable que la collecte de ces données supplémentaires soit particulièrement utile dans les cas où les interventions ne promettent des résultats qu'à relativement long terme. Les entrevues supplémentaires, si elles sont jugées cruciales pour une évaluation, seraient probablement confiées à des entrepreneurs indépendants. Peu importe qu'une entrevue de suivi soit prévue ou non dans le cadre de l'évaluation, le groupe d'experts estime que DRHC devrait prendre des dispositions pour que les participants à l'évaluation puissent être suivis dans le temps, à partir des données administratives relatives à l'AE et (en principe) à la rémunération (voir l'argument suivant).

6. **Il conviendrait d'utiliser des données administratives dans toutes les évaluations pour mesurer certains résultats obtenus par les participants.** Il se peut que des questions de temps empêchent l'utilisation de données administratives sur la rémunération (tirées des feuillets T4) pour mesurer les résultats obtenus par les participants dans le cadre des évaluations telles qu'elles ont été données à contrat actuellement (bien qu'on puisse envisager des contrats de suivi à cet égard), mais il conviendrait d'exploiter le plus possible les données administratives de l'AE. Celles-ci peuvent fournir les mesures les plus exactes des prestations d'assurance-emploi touchées par les participants et elles peuvent également nous éclairer sur la validité des résultats de l'enquête en ce qui concerne l'emploi. On peut également utiliser des données administratives dans le cadre des évaluations pour élaborer des mesures semblables à celles qui ont été élaborées pour le projet des indicateurs à moyen terme (IMT), ce qui faciliterait la comparaison entre les deux études (voir la section E ci-dessous). L'utilisation de données administratives pour mesurer les résultats obtenus par les participants présente également des avantages qui iraient bien au-delà des contrats d'évaluation à titre individuel. En principe, il devrait être possible de suivre les membres du groupe des participants et du groupe témoin pendant de nombreuses années grâce à ces données. Leur utilisation permettrait de déterminer si les répercussions du programme observées lors des évaluations se sont maintenues ou ont disparu rapidement. Il est également possible que les fichiers longitudinaux plus longs que permettraient d'établir les données administratives nous renseignent sur la validité des estimations initiales des répercussions, car ils exploiteraient plus pleinement les variations mesurées dans les caractéristiques de la rémunération des participants et des membres du groupe témoin au fil du temps.

7. **Il conviendrait d'envisager des analyses coûts-avantages et des analyses de rentabilité, mais elles joueront vraisemblablement un rôle secondaire dans les évaluations.** Les estimations de l'impact établies dans le cadre des évaluations pourraient jouer un rôle important en servant de base à des analyses coûts-avantages et à des analyses de rentabilité. Il serait facile de faire une analyse de rentabilité relativement simple, en supposant qu'on dispose de données au sujet des coûts

différentiels des interventions. L'utilité d'une telle analyse dépend beaucoup de la capacité d'estimer fidèlement les répercussions d'interventions particulières — autant du point de vue des modèles quasi expérimentaux que de celui de la détermination de la taille adéquate des échantillons si on veut obtenir des estimations suffisamment détaillées. Il pourrait quand même être possible de faire certaines comparaisons rudimentaires entre les interventions.

La réalisation d'analyses coûts-avantages exhaustives à partir des évaluations pourrait cependant poser des difficultés plus grandes, notamment en raison de l'importance des budgets prévus. Entre autres obstacles majeurs susceptibles de nuire à la tenue d'une analyse coûts-avantages exhaustive, soulignons le fait qu'un grand nombre de répercussions sociales découlant du programme des PEMS pourraient être difficiles à mesurer, qu'il peut s'avérer difficile de prévoir l'impact à long terme des programmes et que, étant donné la taille globale du programme, les effets de déplacement pourraient être significatifs dans certaines régions. Il est particulièrement difficile de trouver des moyens de régler ce dernier problème de façon simple. Pour toutes ces raisons, le groupe d'experts est convaincu que les budgets modestes qui ont été prévus pour les diverses évaluations ne permettraient pas de réaliser les recherches qui s'imposeraient pour faire une analyse coûts-avantages viable. Mais il croit fermement qu'une sorte quelconque d'analyse coûts-avantages devrait constituer le but ultime des évaluations, car les parties intéressées voudront savoir si la valeur des programmes est conforme à leurs coûts. Le groupe croit, par conséquent, qu'il pourrait être prudent que DRHC examine des moyens de mener séparément une analyse coûts-avantages à partir des données combinées[42] de diverses évaluations.

42 À tous le moins, il y aurait lieu d'encourager les fournisseurs à fournir des ensembles de données pour usage public que d'autres chercheurs pourraient combiner pour mettre au point des analyses coûts-avantages ou d'autres types d'analyses.

E. Coordination avec le projet des indicateurs à moyen terme (IMT)

Les évaluations des ententes sur le développement du marché du travail (EDMT) joueront un rôle important dans le processus qui mènera à l'élaboration d'indicateurs à moyen terme pour le programme des PEMS (le « projet des IMT »). Comme ces indicateurs seront élaborés principalement à partir de données administratives, les évaluations donnent l'occasion de déterminer les lacunes éventuelles d'indicateurs fondés seulement sur des données relativement limitées. Une telle comparaison pourra contribuer à déterminer si le projet des IMT réussit à saisir tous les effets (positifs et négatifs) des interventions de prestations d'emploi (PE). Les données plus riches recueillies dans le cadre des évaluations peuvent également représenter des moyens simples d'améliorer les indicateurs à moyen terme prévus (par exemple, si les données de plusieurs sources administratives sont combinées ou si des méthodes novatrices sont utilisées pour regrouper les données au fil du temps). Par ailleurs, le projet des IMT, étant donné que ses échantillons et ses activités récurrentes seront beaucoup plus vastes, permettra d'examiner les effets sur divers sous-groupes et les innovations récentes apportées au programme, qu'il serait impossible d'examiner dans le cadre des évaluations de EDMT. Les effets des IMT dans le cadre de telles analyses seront plus utiles si les évaluations ont permis d'établir des mesures de validité de base montrant les rapports entre ces effets estimés et les mesures plus détaillées des résultats provenant des évaluations de EDMT. Il sera donc essentiel que les deux projets soient coordonnés. Aux fins de cette coordination, le groupe d'experts a formulé les recommandations suivantes :

1. **Les échantillons des évaluations devraient être sélectionnés de façon à faciliter la comparaison avec les résultats du projet des IMT**. Un grand nombre des suggestions faites dans les sections A et B ci-dessus visent cet objectif. En général, on espère que les échantillons des évaluations pourront être considérés comme des échantillons aléatoires des populations plus vastes qui seraient utilisées pour l'élaboration des IMT pendant une période donnée. De cette façon, les résultats des évaluations permettraient de faire des analyses directes des IMT. Il faudra tenir des consultations pendant les phases d'élaboration des évaluations des EDMT pour veiller à ce que cet objectif soit respecté.

2. **Les entrepreneurs chargés des évaluations de EDMT devraient, lorsque c'est possible, élaborer des mesures semblables aux IMT pour leurs échantillons (ou faciliter la tâche d'autres chercheurs qui s'en chargeront).** Pour examiner l'efficacité des indicateurs à moyen terme et la façon dont ils pourraient être améliorés, il faudra utiliser des microdonnées tirées des évaluations. Cela ne sera toutefois possible que si l'on s'est donné la peine d'intégrer les données administratives appropriées aux fichiers de recherche. De toute évidence, certaines données qui serviront à l'élaboration des IMT ne seront peut-être pas accessibles aux évaluateurs dans les délais prévus (ce sera probablement le cas des données plus

récentes de l'Agence des douanes et du revenu du Canada (ADRC). Dans ces circonstances, les fichiers de recherche devraient être conçus de telle façon que ces données puissent y être intégrées ultérieurement.

3. **On devrait s'efforcer de coordonner les évaluations prévues et le processus d'élaboration des IMT.** On risque de rater un grand nombre des possibilités de coordination entre les évaluations des EDMT et le projet des IMT si les deux projets ne sont pas menés approximativement en même temps. Puisque les dates de réalisation des évaluations sont plus ou moins déterminées en vertu des EDMT, les décideurs qui s'intéressent à l'élaboration des IMT devraient, lorsque c'est possible, s'inspirer du même calendrier.

F. Sommaire des questions sur lesquelles il faudra revenir au moment de la mise en œuvre

Ce document présente un cadre générique de pointe pour les évaluations des prestations d'emploi et mesures de soutien (PEMS). Au moment de mettre en œuvre les méthodes qui y sont exposées, les réalisateurs des évaluations devraient tenir compte de la situation particulière dans la région visée. Ils pourront ainsi justifier toute hypothèse reposant sur le cadre, de même que les modifications et les extensions. Celles-ci devraient être bien étayées et rationalisées en fonction des constats, des raisons qui justifient l'application des techniques, de même que des connaissances et des théories sous-jacentes. Les résultats seront d'autant plus crédibles que plusieurs techniques auront été utilisées simultanément, selon la méthode des recoupements. De surcroît, pour tenir compte de certaines des incertitudes inhérentes aux modèles quasi expérimentaux, il conviendrait d'encourager les réalisateurs à fournir des fourchettes d'estimations, plutôt que des estimations monovalentes.

On a abordé ici de façon très générale certains des grands enjeux qui se manifesteront probablement au fur et à mesure de la conception des évaluations des PEMS. Notre objectif n'était pas de fournir un plan explicite pour chacune de ces évaluations, mais plutôt de décrire certaines des questions méthodologiques universelles sur lesquelles il faudra se pencher. Le groupe d'experts est convaincu que les évaluations des PEMS donnent l'occasion de faire des travaux de recherche de pointe en matière d'évaluation, tout en recueillant une information substantielle qui sera à la fois valable et utile pour les décideurs.

Le groupe d'experts a déterminé diverses questions sur lesquelles il faudra se pencher de plus près. C'est aux réalisateurs qu'il reviendra de s'occuper des détails.

1. **La répartition des tâches entre les diverses évaluations devrait se révéler profitable.** Comme on ne peut pas s'attendre à ce que toutes les évaluations fassent « tout », le groupe d'experts pense qu'une certaine répartition des tâches entre les évaluations serait appropriée et qu'elle permettrait d'obtenir les meilleurs résultats globaux. Voici trois exemples de répartition des tâches : 1) une analyse plus approfondie de certaines interventions (p. ex., le travail indépendant) dans certaines évaluations; 2) certaines évaluations pourraient mettre davantage l'accent sur les résultats de non emploi plus difficiles à mesurer; 3) si une évaluation donne l'occasion de mettre en œuvre une répartition au hasard, il conviendrait de l'exploiter.

2. **Il conviendrait d'élaborer des procédures permettant de déterminer la qualité des résultats des évaluations et de les faire connaître.** Même si certaines personnes qui ne font pas partie de la collectivité des chercheurs seront convaincues que *les conclusions* des évaluations des PEMS ne comporteront aucune surprise pour elles,

il semble beaucoup plus probable que les résultats réels soient assez différents d'une région à l'autre. Une partie de cette variation pourra être attribuable à des différences véritables dans les répercussions du programme, une autre partie à des différences dans la conception et la mise en œuvre du programme, une autre encore à des différences dans la conception des évaluations, et à des différences dans la demande de main-d'œuvre à l'échelle des régions et enfin, une autre partie à des facteurs purement aléatoires (p. ex., des différences au chapitre du marché du travail). Le groupe d'experts pense que si la communication de ces variations ne se fait pas de façon ordonnée, c'est la validité globale perçue des évaluations qui risque d'en souffrir. Pour atténuer de tels problèmes, il semblerait souhaitable que d'entrée de jeu, on mette sur pied des procédures de communication des résultats des évaluations selon lesquelles ces résultats seraient replacés dans un contexte global. Ce contexte exposerait les aspects particuliers de chaque évaluation à titre individuel, tout en expliquant comment ces résultats correspondent à un profil plus global.

3. **Il conviendrait d'envisager des moyens de mettre en commun les résultats des évaluations dans les analyses subséquentes.** Le groupe d'experts pense que la mise en commun des résultats des évaluations, même si elle peut poser des problèmes de non-comparabilité statistique, devrait faire partie intégrante du processus d'évaluation des PEMS, et que les fournisseurs devraient être encouragés à rendre leurs données disponibles à de telles fins. On y retrouverait divers avantages, notamment les suivants :

    - des échantillons plus vastes, de sorte que la précision des estimations s'en trouverait améliorée;
    - la capacité d'étudier les effets de différences éventuelles dans le contenu des programmes;
    - des indications sur le rendement des interventions de prestations d'emploi (PE) dans différents marchés du travail;
    - de meilleures possibilités d'examiner les résultats obtenus par certains sous-groupes de participants;
    - la possibilité d'explorer des enjeux qui vont peut-être au-delà du mandat des évaluations à titre individuel — par exemple entreprendre une analyse de rentabilité détaillée du programme dans son ensemble.

Pour atteindre ces objectifs importants, il sera nécessaire d'imposer un certain degré de comparabilité entre les évaluations des points de vue de la collecte et de l'accessibilité des données. Des normes en ce sens doivent être prévues dès le début — il se révélera peut-être impossible de mettre en commun les données de plusieurs évaluations si chaque évaluateur a toute latitude pour choisir ses propres solutions aux problèmes concernant les données.

4. **Il conviendrait d'envisager, lorsque c'est possible, d'utiliser des instruments communs de collecte des données dans toutes les évaluations.** Cette conclusion dérive des trois arguments précédents. Les problèmes relatifs à la mise en commun et à la communication des résultats des évaluations seront tous exacerbés si chaque évaluation a ses propres instruments de collecte des données. Le groupe d'experts est donc fermement convaincu que certains modules communs (particulièrement en ce qui concerne les antécédents sur le marché du travail et les résultats obtenus par les participants) devraient être intégrés à toutes les évaluations. Pour élaborer de tels modules, il faudra prévoir des consultations exhaustives entre divers intervenants du programme des PEMS, à la fois pour veiller à ce que l'instrument de collecte des données soit de qualité élevée et pour faire en sorte que des données soient recueillies sur tous les résultats jugés pertinents du point de vue des politiques publiques (y compris divers résultats socioéconomiques qui n'ont pas été examinés explicitement ici).

5. **Quel degré d'appariement pourra-t-on faire avant le tirage des échantillons?** Comme on l'a vu dans les sections B et C, il est impossible de concevoir un plan d'échantillonnage pour les évaluations tant qu'on ne connaît pas le degré d'appariement qui pourra être fait à partir des données administratives. C'est pourquoi le groupe d'experts est convaincu qu'il est important d'entreprendre certains travaux préliminaires à partir des données administratives, pour déterminer les limites du processus d'appariement. En général, on peut s'attendre à ce que ces limites soient plus importantes s'il est impossible d'utiliser les données de l'agence des douanes et du revenu du canada (ADRC) pour le tirage des échantillons dans les premières évaluations. Il sera essentiel que les lacunes que représentera l'utilisation des seules données de l'AE pour choisir les groupes témoins soient étayées au moment de l'application de la méthodologie et que des moyens de combler ces lacunes soient proposés dans le cadre de l'enquête. Il sera particulièrement important de déterminer dans quelle mesure l'incapacité d'établir de bons appariements à partir des seules données de l'AE influencera la répartition de l'échantillon entre les groupes de participants et les groupes témoins.

6. **Quel rôle les estimations VI devraient-elles jouer dans les évaluations?** L'utilisation d'estimateurs VI soulève certaines des questions les plus difficiles en ce qui concerne les évaluations. En effet, la détermination des situations dans lesquelles de tels estimateurs donneraient des estimations cohérentes des répercussions est souvent ambiguë; en outre, la détermination de la qualité des estimateurs relève autant de l'art que de la science. Le groupe d'experts croit donc fermement que le rôle des estimateurs VI dans les évaluations devrait être examiné. Voici quelques-unes des questions auxquelles il conviendrait de s'intéresser en ce qui concerne l'utilisation de tels estimateurs :

    - Quelles variables seront utilisées aux fins de l'identification que nécessitent les estimateurs VI? Comment ces données seront-elles recueillies? Comment les restrictions de l'identification seront-elles testées?

- Comment les propriétés des estimateurs VI seront-elles influencées par la façon dont les échantillons des participants et des témoins seront sélectionnés? Une analyse de simulation peut-elle permettre de comprendre la relation entre l'estimation VI et l'appariement de l'échantillon?

- Comment la robustesse des estimateurs VI peut-elle être évaluée? Comment devrait-on procéder pour déclarer et évaluer les autres spécifications possibles pour les estimateurs VI?

7. **Il conviendrait d'examiner plus à fond la question de la taille des échantillons pour chaque application.** Selon les calculs présentés ici, on aurait besoin d'échantillons d'au moins 2 000 participants et 2 000 témoins pour obtenir que les estimations seront suffisamment précises pour déterminer les impacts pertinents des politiques. Mais cette estimation de précision s'est faite à partir d'une information très limitée ou uniquement à partir de comparaisons par rapport à d'autre études. De toute évidence, un examen plus minutieux de la question propre à chaque évaluation s'impose avant que d'importantes ressources soient engagées. Un tel examen devrait porter sur des questions comme les suivantes : 1) le désir d'estimer les impacts d'interventions particulières; 2) comment le groupe témoin peut être « partagé » entre différentes interventions, s'il peut l'être du tout; 3) comment les méthodes d'analyse se répercuteront vraisemblablement sur la taille des échantillons disponibles; et 4) comment la probabilité de non-réponse influence le calcul de la taille des échantillons.

Bibliographie

ASHENFELTER, Orley. « Estimating the Effects of Training Programs on Earnings », *Review of Economics and Statistics*, janvier 1978, p. 47-57.

ASHENFELTER, Orley. « Estimating the Effect of Training Programs on Earnings with Longitudinal Data », dans *Evaluating Manpower Training Programs*, ouvrage collectif publié sous la direction de F. Bloch, Greenwich, CT, JAI Press, 1979, p. 97-117.

ASHENFELTER, Orley, et David CARD. « Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs », *Review of Economics and Statistics*, juin 1985, p. 648-660.

DEHEJIA, Rajeev, et Sadek WAHBA. « Propensity Score Matching Methods for Nonexperimental Causal Studies », National Bureau of Economic Research, Cambridge, MA, Working Paper No. 6829, 1998.

DEHEJIA, Rajeev, et Sadek WAHBA. « Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs », *Journal of the American Statistical Association*, décembre 1999, 94(448), p. 1053-62.

GREENBERG, David, et Mark Shroder. « The Digest of Social Experiments, Second Edition », Washington, D.C., The Urban Insitute Press, 1997.

HECKMAN, James. « Varieties of Selection Bias », *American Economic Review Papers and Proceedings*, mai 1990, 80(2), p. 313-18.

HECKMAN, James, et Joseph HOTZ. « Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training », *Journal of the American Statistical Association*, décembre 1989, 84(408), p. 862-74.

HECKMAN, James, Hidehiko ICHIMURA, Jeffrey SMITH et Petra TODD. « Characterizing Selection Bias Using Experimental Data », *Econometrica*, septembre 1998a, 66(5), p. 1017-98.

HECKMAN, James, Hidehiko ICHIMURA et Petra TODD. « Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme », *Review of Economic Studies*, octobre 1997, 64(4), p. 605-54.

HECKMAN, James, Hidehiko ICHIMURA et Petra TODD. « Matching as an Econometric Evaluation Estimator », *Review of Economic Studies*, avril 1998b, 65(2), p. 261-94.

HECKMAN, James, Robert LALONDE et Jeffrey SMITH. « The Economics and Econometrics of Active Labor Market Programs », dans *Handbook of labor economics*, vol. 3A, ouvrage collectif publié sous la direction de Orley Ashenfelter et David Card, Amsterdam, North-Holland, 1999, p. 1865-2097.

HECKMAN, James, et Jeffrey SMITH. « The Preprogramme Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies », *Economic Journal*, juillet 1999, 109(457), p. 313-48.

ICHIMURA, Hidehiko, et Christopher TABER. « Direct Estimation of Policy Impacts », National Bureau of Economic Research, Cambridge, MA, Technical Working Paper No. 254, 2000.

IMBENS, Guido, et Joshua ANGRIST. « Identification and Estimation of Local Average Treatment Effects », *Econometrica*, mars 1994, 62(2), p. 467-75.

LALONDE, Robert. « Evaluating the Econometric Evaluations of Training Programs with Experimental Data », *American Economic Review*, septembre 1986, 76(4), p. 604-20.

NEWEY, Whitney, et James POWELL. « Instrumental Variables for Nonparametric Models », manuscrit inédit, Université Princeton, 1989.

NICHOLSON, Walter. « Assessing the Feasibility of Measuring Medium Term Net Impacts of the EBSM Program in Nova Scotia », document de travail rédigé pour DRHC, mars 2000.

NICHOLSON, Walter. « Problèmes de conception des évaluations sommatives : Sommaire des discussions… », DRHC, mars 2001.

ROSENBAUM, Paul, et Donald RUBIN. « The Central Role of the Propensity Score in Observational Studies for Causal Effects », *Biometrika*, avril 1983, 70(1), p. 41-55.

SMITH, Jeffrey, et Petra TODD. « Reconciling Conflicting Evidence on Performance of Propensity Score Matching Methods », *American Economic Review Papers and Proceedings*, mai 2001, 91(2), p. 112-18.

Annexe A[43] *: Évaluations sommatives des prestations d'emploi et mesures de soutien ainsi que des ententes sur le développement du marché du travail — répercussions éventuelles des programmes à examiner/mesurer*

La présente annexe expose un éventail de répercussions éventuelles des programmes. Cette « démarche des répercussions éventuelles des programmes » est conforme au cadre qu'utilise depuis quelque temps le Bureau du vérificateur général pour examiner la mesure de l'efficacité des programmes menée par les ministères et organismes fédéraux.

Aux fins de l'évaluation, il *ne s'agit pas* de mesurer les résultats individuels par rapport à un objectif quantitatif ou à un niveau-repère quantitatif précis à atteindre. La démarche consiste plutôt à examiner et à mesurer, de la manière la plus complète possible, ce que les programmes financés dans le cadre des (PEMS)/(EDMT) ont accompli (leurs activités) et ce qu'ont été les résultats.

Il s'agit ici d'adopter la même démarche d'évaluation de base dans toutes les régions, puis d'élaborer une méthode uniforme à l'échelle nationale pour l'analyse agrégée des résultats globaux des PEMS/EDMT. Par ailleurs, une marge de manœuvre sera ménagée au niveau local en ce qui concerne les thèmes qui seront abordés dans les évaluations; ainsi, les régions pourront décider d'y intégrer d'autres questions/indicateurs, à leur choix.

Il est présumé que les consultants retenus rédigeront, une fois le processus d'évaluation amorcé, des rapports de méthodologie abordant plus à fond le processus de mesure dans les secteurs choisis.

---

43 La présente annexe ne fait pas partie du cadre élaboré par le groupe d'experts. Leurs débats ont porté en grande partie sur la méthodologie expliquée dans le rapport. Les discussions au sujet d'éléments bien précis à mesurer ont été très limitées. La clarification des problèmes de mesures pourrait constituer une lourde tâche future pour les membres du groupe d'experts.

La section intitulée « Répercussions éventuelles du programme… », plus bas, énumère des questions considérées importantes dans plusieurs politiques liées à l'emploi et dans des études menées par différents chercheurs et gouvernements, à l'échelle nationale ou internationale. En ce qui concerne les évaluations des PEMS, les questions peuvent être tirées des éléments suivants:

- les objectifs des EDMT/PEMS tels qu'on les trouve dans la législation sur l'assurance-emploi (AE) et les ententes relatives au marché du travail;
- les questions soulevées dans le cadre d'évaluations, d'analyses de politique ou bien d'énoncés de politique en rapport avec l'emploi;
- les questions soulevées et examinées dans le cadre des évaluations formatives des EDMT/PEMS;
- les questions soulevées lors de la planification préliminaire des évaluations sommatives et dans le cadre du projet-pilote des indicateurs à moyen terme;
- les répercussions éventuelles que des investissements de cette nature peuvent avoir sur le marché du travail et l'économie en général;
- les tendances socioéconomiques émergentes au Canada et chez ses partenaires commerciaux.

Comme il en est question dans le rapport, la méthodologie comprend diverses techniques de mesure qui peuvent être appliquées à un vaste éventail de résultats ou d'indicateurs. Par conséquent, le cas échéant, Développement des ressources humaines canada (DRHC) s'attendra à ce que les résultats soient mesurés comme suit : résultats obtenus par les participants après le programme, différences entre les expériences avant et après le programme, et comparaisons des expériences des participants et des non-participants.

RÉPERCUSSIONS ÉVENTUELLES DU PROGRAMME
RÉSULTATS OBTENUS PAR LES PARTICIPANTS

CLIENTS QUI ONT REÇU DE L'AIDE À TITRE INDIVIDUEL

- POURCENTAGE DES PARTICIPANTS QUI ONT TROUVÉ DES EMPLOIS DANS LA PÉRIODE QUI A SUIVI LE PROGRAMME

- POURCENTAGE DES SEMAINES DE TRAVAIL DANS LA PÉRIODE QUI A SUIVI LE PROGRAMME

- DURÉE MOYENNE DES PÉRIODES D'EMPLOI

- NIVEAUX MOYENS DES PRESTATIONS D'AE/D'AIDE SOCIALE TOUCHÉES

- POURCENTAGE DES SEMAINES DE PRESTATIONS D'AE/D'AIDE SOCIALE DANS LA PÉRIODE QUI A SUIVI LE PROGRAMME (autosuffisance des travailleurs — réduction de la dépendance envers l'AE. et l'aide sociale)

- RECOURS À L'AE PAR RAPPORT À L'AIDE SOCIALE AVANT ET APRÈS LE PROGRAMME

- RÉMUNÉRATION MOYENNE PAR RAPPORT AU SEUIL DE FAIBLE REVENU (ou autres mesures de la rémunération ou du revenu utilisées sur le marché ou au sein de la collectivité)

- REVENU MOYEN PAR RAPPORT AU SEUIL DE FAIBLE REVENU (ou autres mesures de la rémunération ou du revenu utilisées sur le marché ou au sein de la collectivité)

- INDICATEURS DE LA QUALITÉ DE VIE (ET SURTOUT LES AMÉLIORATIONS QUE CERTAINS PARTICIPANTS AFFICHENT PAR RAPPORT À CEUX QUI N'ONT CONNU AUCUN CHANGEMENT : MOTIVATION; SATISFACTION FACE À LA VIE). (Nota : c'est pour les raisons suivantes que cette question est intégrée à la démarche proposée pour l'évaluation : a) parce qu'il s'agit de l'un des objectifs globaux officiels du ministère; et b) parce que dans les ouvrages techniques publiés sur la question en général [y compris les ouvrages d'autres pays], de tels effets peuvent être associés à des programmes du genre de ceux qui sont prévus dans les EDMT)

- MOBILITÉ GÉOGRAPHIQUE/RÉINSTALLATION DES CLIENTS

PRINCIPALES CARACTÉRISTIQUES DE LA CLIENTÈLE (et surtout les principaux genres de problèmes auxquels font face les clients qui reçoivent de l'aide en vertu des EDMT).

- ÂGE/SEXE PAR RAPPORT À L'ENSEMBLE DES CHÔMEURS (CHÔMAGE CHEZ LES JEUNES / ENJEUX DE L'ÉQUITÉ ENTRE LES HOMMES ET LES FEMMES)

- PROPORTION DES CLIENTS QUI ÉPROUVENT DES PROBLÈMES CHRONIQUES D'EMPLOI/QUI SONT AUX PRISES AVEC DES OBSTACLES À L'EMPLOI AVANT LA PARTICIPATION AU PROGRAMME (d'après une recension de la littérature sur les principaux genres d'obstacles à l'emploi)

- PROPORTION DES CLIENTS DONT ON A DÉTERMINÉ QU'ILS ÉTAIENT DES TRAVAILLEURS DÉPLACÉS AVANT DE RECEVOIR L'AIDE DU PROGRAMME

- NIVEAUX DE SCOLARITÉ (ENJEUX DE L'ALBHABÉTISATION FONCTIONNELLE)

- CATÉGORIES PROFESSIONNELLES ET NIVEAUX DE QUALIFICATION AVANT ET APRÈS LE PROGRAMME (lacunes dans les compétences/acquisition de compétences)

- APPRENTIS (MATÉRIEL DESCRIPTIF SEULEMENT)

COLLECTIVITÉS LOCALES

- INITIATIVES DE DÉVELOPPEMENT DU MARCHÉ DU TRAVAIL POUR AIDER LES COLLECTIVITÉS (PAR. EX., AIDE AUX ORGANISMES LOCAUX DE DÉVELOPPEMENT ÉCONOMIQUE/EMPLACEMENT DES INDUSTRIES) (MATÉRIEL DESCRIPTIF SEULEMENT)

GOUVERNEMENTS

- RAPPORT COÛT-EFFICACITÉ DES INTERVENTIONS (par ex., améliorations nettes de l'employabilité et de la durée d'emploi parmi diverses catégories de clients par rapport aux coûts de l'aide fournie dans le cadre du programme)

- LES PEMS/EDMT COMME MÉCANISME D'OFFRE DE MAIN-D'ŒUVRE (évaluation de l'ampleur des activités mises en œuvres dans le cadre des EDMT par rapport aux besoins de l'économie en matière d'expansion et de remplacement des effectifs)

- ÉCONOMIES BRUTES POUR LE COMPTE DE L'AE DÉCOULANT DES PEMS/EDMT

- ÉCONOMIES NETTES POUR LE COMPTE DE L'AE DÉCOULANT DES PEMS/EDMT

SUIVI DES ÉTUDES FORMATIVES

- CERTAINES QUESTIONS SOULEVÉES DANS LE CADRE DES ÉTUDES FORMATIVES

- CERTAINES QUESTIONS QUI ONT ÉMERGÉ DEPUIS LA FIN DES ÉTUDES FORMATIVES